An empirical analysis of journal policy effectiveness for computational reproducibility

Victoria Stodden[a,1], Jennifer Seiler[b], and Zhaokun Ma[b]

[a]School of Information Sciences, University of Illinois at Urbana–Champaign, Champaign, IL 61820; and [b]Department of Statistics, Columbia University, New York, NY 10027

Edited by David B. Allison, Indiana University Bloomington, Bloomington, IN, and accepted by Editorial Board Member Susan T. Fiske January 9, 2018 (received for review July 11, 2017)

A key component of scientific communication is sufficient information for other researchers in the field to reproduce published findings. For computational and data-enabled research, this has often been interpreted to mean making available the raw data from which results were generated, the computer code that generated the findings, and any additional information needed such as workflows and input parameters. Many journals are revising author guidelines to include data and code availability. This work evaluates the effectiveness of journal policy that requires the data and code necessary for reproducibility be made available postpublication by the authors upon request. We assess the effectiveness of such a policy by (*i*) requesting data and code from authors and (*ii*) attempting replication of the published findings. We chose a random sample of 204 scientific papers published in the journal *Science* after the implementation of their policy in February 2011. We found that we were able to obtain artifacts from 44% of our sample and were able to reproduce the findings for 26%. We find this policy—author remission of data and code postpublication upon request—an improvement over no policy, but currently insufficient for reproducibility.

reproducible research | data access | code access | reproducibility policy | open science

The creation of digital scholarly artifacts such as datasets and code is an integral part of computational research, and a broad movement has emerged to encourage the dissemination of artifacts that underlie published results (1–6). A shift is occurring in the research community toward the routine dissemination of digital artifacts, including journal publication requirements that authors make available data and code sufficient for replication purposes upon request after publication (7). Such a policy was implemented by *Science* on February 11, 2011 (8, 9):

> All data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of Science. All computer codes involved in the creation or analysis of data must also be available to any reader of Science. After publication, all reasonable requests for data and materials must be fulfilled. Any restrictions on the availability of data, codes, or materials, including fees and original data obtained from other sources (Materials Transfer Agreements), must be disclosed to the editors upon submission...

Science suggests using established community repositories to host data. If that is not possible, the policy specifies the use of the supplemental materials section associated with the publication, or, failing that, remitting it to *Science* and posting it on an institutional website where it will be accessible for at least 5 y (9).

Although the policy explicitly states that code must be shared as well as data, the policy does not suggest specific repositories or give instructions for hosting and sharing code and computational methods, as they do for data. Sharing code is perhaps not so straightforward, as there is no consensus regarding repositories, metadata, or computational provenance, as there is for data sharing in many disciplines (e.g., ref. 10).

In this work, we seek to test the efficacy of the *Science* policy in bringing about data and code availability, as well as computational reproducibility of published results. We use a survey instrument to test the availability of data and code for articles published in *Science* in 2011–2012. We then use the scientific communication standards from the 2012 Institute for Computational and Experimental Research in Mathematics (ICERM) workshop report to evaluate the reproducibility of articles for which artifacts were made available (11). We then assess the impact of the policy change directly, by examining articles published in *Science* in 2009–2010 and comparing artifact ability to our postpolicy sample from 2011–2012. Finally, we discuss possible improvements to journal policies for enabling reproducible computational research in light of our results.

Results

We emailed corresponding authors in our sample to request the data and code associated with their articles and attempted to replicate the findings from a randomly chosen subset of the articles for which we received artifacts. We estimate the artifact recovery rate to be 44% with a 95% bootstrap confidence interval of the proportion $[0.36, 0.50]$, and we estimate the replication rate to be 26% with a 95% bootstrap confidence interval $[0.20, 0.32]$.

Procuring Data and Code. Our sample comprised 204 computational articles that appeared in *Science* magazine in 2011–2012 (see Methods for details). For the purposes of this study, we deemed a computational publication one whose findings relied on the use of computational and data-enabled methods (12). Twenty-four of these articles contained sufficient information (via links or in the supporting information) for us to locate the artifacts without contacting the authors. We emailed the remaining 180 authors requesting the data and code used to generate the results in their publication. A total of 131 of the authors replied to our request, and 3 emails bounced.

At least some of the requested material was provided by 36% of the 180 emailed authors (Table 1). We found that 11% were unwilling to provide the data or code without further information regarding our intentions, and 11% asked us to contact someone

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Reproducibility of Research: Issues and Proposed Remedies," held March 8-10, 2017, at the National Academy of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Reproducibility.

Author contributions: V.S. designed research; V.S., J.S., and Z.M. performed research; V.S., J.S., and Z.M. analyzed data; and V.S. and J.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. D.B.A. is a guest editor invited by the Editorial Board.

Data deposition: We have created a repository at https://github.com/ReproducibilityIn-Publishing/Science-2018 containing supplemental material, including the code and email templates that were used in the survey.

[1]To whom correspondence should be addressed. Email: victoria@stodden.net.

Published online March 12, 2018.

Table 1. Responses to emailed requests (n = 180)

Type of response	Count	Percent, %
Did not share data or code:		
Contact another person	20	11
Asked for reasons	20	11
Refusal to share	12	7
Directed back to supplement	6	3
Unfulfilled promise to follow up	5	3
Impossible to share	3	2
Shared data and code	65	36
Email bounced	3	2
No response	46	26

else who worked on the article, six of whom were copied by the corresponding author with no further response. We found that 7% refused to share data and/or code, and 2% gave reasons they could not ethically share or had size or other sharing limitations. Each response was classified into one category only, according to their principal concern. Responses tended to focus on a single barrier, making the categorization straightforward. Some examples of the responses we received are included below.

As Table 1 shows, this policy procures data/code for 65 of the 180 emailed articles, or 36% of this sample. This gives a total of 89 articles in our sample for which we had artifacts, including the 24 which contained sufficient information.

For these 89 articles, we evaluated by inspection whether it appeared possible to carry out a replication of the published results and judged that 56 were potentially reproducible with our resources. If we did not have time and computational resource constraints, we judged that we could have included 9 more. Additional articles may have been reproducible with further interaction with the authors.

There appeared to be some confusion among authors, some of whom seemed to be unaware of *Science*'s data and code sharing requirement. We can most easily demonstrate this with some anonymized author responses that highlight some of the barriers to sharing they perceived:

> When you approach a PI for the source codes and raw data, you better explain who you are, whom you work for, why you need the data and what you are going to do with it.

> I have to say that this is a very unusual request without any explanation! Please ask your supervisor to send me an email with a detailed, and I mean detailed, explanation.

> The data files remains our property and are not deposited for free access. Please, let me know the purpose you want to get the file and we will see how we can help you.

> We do not typically share our internal data or code with people outside our collaboration.

> The code we wrote is the accumulated product of years of effort by [redacted] and myself. Also, the data we processed was collected painstakingly over a long period by collaborators, and so we will need to ask permission from them too.

> Normally we do not provide this kind of information to people we do not know. It might be that you want to check the data analysis, and that might be of some use to us, but only if you publish your findings while properly referring to us.

> Thank you for your interest in our paper. For the [redacted] calculations I used my own code, and there is no public version of this code, which could be downloaded. Since this code is not very user-friendly and is under constant development I prefer not to share this code.

> I'm sorry, but our computer code was not written with an eye toward distributing for other people to use. The codes are not documented and we don't have the time or resources to document them. If you have a particular calculation you would like done and it is not a major extension of what we are presently set up to do, we might be able to run the codes for you.

> R is a free software package available at www.r-project.org/ I used R for the [redacted] models. As you probably know, [redacted] and [redacted] are quite complicated. But I don't have to tell you that given that you are a statistics student! I used Matlab for the geometry.

These responses can be contrasted with replies from authors who were not only willing to share, but had clearly made an effort to make their methods accessible and well documented:

> Our program [redacted] is available here [URL redacted] (documentation and tutorials were included)

> If you go to [URL redacted], under the publications, I have a link to the gitHub repository. I don't know if I have all of the raw simulated data, but I certainly have the processed data used to make the plots. What do you need? All of the simulated data could of course be regenerated from the code.

> Please find attached a .zip file called [redacted].zip that has the custom MATLAB [redacted] analysis code. If you run Masterrunfigureone.m this will generate several panels from the paper.

> In the next email I will enclose the custom image analysis software. This can also be accessed from [URL redacted] where there is a manual and tutorial.

> Please let me know if you have any troubles, or if there is anything else I can help with.

We regret being unable to reveal these authors' website and repository information due to our confidentiality restrictions, as there were some very complete and excellent examples of how to publish reproducible research. While some authors who provided publicly accessible data, code, and documentation made use of resources such as sourceforge.com and GitHub.com, many more simply had links to university ftp locations or created barebones websites containing lists of files.

Our next step is to attempt replication on a random sample of the 56 articles we judged potentially reproducible.

Reproducing Published Results. We randomly chose 22 articles from the 56 deemed likely to be reproducible, and we were able to replicate the results in the publication for all but 1. The one that was deemed irreproducible used a large community dataset and provided links to software for data extraction tools which were no longer usable or available.

As the papers were randomly selected from among those carefully chosen to be likely reproducible with our resources, missing data or code in these 22 papers was rare. The issues most commonly seen in the remaining articles we deemed unlikely to reproduce were missing scripts, documentation, or parameters. Few papers cited visualization tools, even when the visualizations in their article were instrumental to support their conclusions. We made the decision to overlook the lack of details regarding the visualization step and considered these papers reproducible if otherwise complete.

It is important to note that the failure to cite both visualization tools as well as common software packages (such as MATLAB) was a widespread failure of the majority of the 204 papers (at least 139 papers failed to cite). It is also important to note that much of the code was received by us via email well after the publication date, and had typically been modified since it had been used to generate the results in the publication, also causing difficulties in replication.

Table 2. ICERM implementation criteria for articles deemed likely to reproduce (n = 56)

ICERM criteria	Percent compliant, %
A precise statement of assertions to be made in the paper.	100
Full statement (or valid summary) of experimental results.	100
Salient details of data reduction & statistical analysis methods.	91
Necessary run parameters were given.	86
A statement of the computational approach, and why it constitutes a rigorous test of the hypothesized assertions.	8
Complete statements of, or references to, every algorithm used, and salient details of auxiliary software (both research and commercial software) used in the computation.	80
Discussion of the adequacy of parameters such as precision level and grid resolution.	79
Proper citation of all code and data used, including that generated by the authors.	79
Availability of computer code, input and output data, with some reasonable level of documentation.	77
Avenues of exploration examined throughout development, including information about negative findings.	68
Instructions for repeating computational experiments described in the article.	63
Precise functions were given, with settings.	41
Salient details of the test environment, including hardware, system software, and number of processors used.	13

Evaluating Current Practices. We checked the ICERM Implementation Criteria (appendix D in the report) for the 56 potentially reproducible papers, grouping them into two sets of results: Implementation information is provided in Table 2, and data and code accessibility are in Table 3 (11, 13).

We assessed the ICERM implementation criteria for all 56 of the articles judged to be potentially reproducible by a thorough reading of the article, supplemental materials, and any provided artifacts.

Even though all 204 papers had at least some computational components, a statement of the computational approach was more rare. A total of 46 of the 56 potentially reproducible papers evaluated contained statements of the computational approach; however, only 7 of the 56 papers mentioned hardware or environmental settings. We found that 86%, or 48 articles, provided the necessary parameters, and 44 of those discussed those parameter choices, although those choices were rarely listed as part of computational instructions. Parameter choices were mostly established via a reading of the analysis and methods and were often distributed throughout the article. A total of 35 papers gave specific instructions for repeating the computational experiments, making reproduction attempts much easier for a researcher.

The next subset of ICERM criteria we applied to the 56 potentially reproducible articles referred to the documentation, archiving, and curation of data and code, and is summarized in Table 3.

Science's guidelines suggest that researchers reference their data deposition site in their acknowledgement section: "We will also ask authors to provide a specific statement regarding the availability and curation of data as part of their acknowledgements" (9).

However only 39 of the 56 articles (70%) did so. Note that only 66% of the 56 articles we deemed to be potentially reproducible provided artifact licensing information. This can be a major stumbling block to reuse and is easily rectifiable (14). Only a little more than half this subsample had openly available code or adequate documentation. Table 3 documents shortcomings in reusability and persistence for digital artifacts.

Reproducing Results. We developed a system to categorize the reproduction efforts, given in Table 4. The most common obstacle we found was missing essential parameters or scripts. Several of the email responses mentioned that they did not keep the small scripts they used to manage their analysis or simulations. If an error were contained in these scripts, we would not be able to identify its exact location or debug it.

There were three additional issues encountered that hindered replication. First, specialized plotting or visualization software was rarely cited or listed anywhere in the article or supplemental materials. Where this happened, for the most part, we relied exclusively on quantitative analysis to verify article conclusions and did not attempt to recreate the relevant figures. Additionally, common packages were rarely cited, although this was often easily deduced. Second, hardware and environmental settings were rarely discussed, although this was often relevant for reproducibility. Last, function calls and well-documented workflows were rare. This meant that function calls and the order of execution had to be deduced from the text, and sometimes by trial and error.

Three articles provided upfront all necessary documentation, scripting, references, and parameters required to replicate the procedures in the papers without adding unnecessary effort for the reader. Six articles had only a single minor oversight each that was easily overcome and did not prevent replication.

The classification given in Table 4 extends previous evaluations of reproducibility levels. Reproducibility was attempted in 2008 for articles published in a multidisciplinary genetics journal, and replication standards involved checking data availability and the match between the data annotation and the published analyses (15). They found they were able to reproduce the figures for 10 of 16 articles. In 2012, replication was attempted for 23 articles that used a specialized software, and the authors documented missing input parameters and missing data as causes of failures to replicate (16). In 2015, a replication analysis was carried out for 67 articles appearing in 13 economics journals, successfully replicating 29 (17). In this work, sources of failure are given as missing data or code, incorrect data or code, missing software, or proprietary data.

Impact due to Policy Change. To evaluate the effectiveness of the 2011 requirements by *Science* magazine, we compared data and

Table 3. ICERM archiving criteria for articles deemed likely to reproduce (n = 56)

ICERM criteria	Percent compliant, %
Data documented to clearly explain what each part represents.	91
Data archived with significant longevity expected.	82
Data location provided in the acknowledgements.	70
Authors have documented use and licensing rights.	66
Software documented well enough to run it and what it ought to do.	57
The code is publicly available with no download requirements.	54
There was some method to track changes/to the software, as well as some certainty that the code is securely archived.	50

Classification	Percent, %
Impossible to reproduce (missing essential code, data, or methodology)	5
Nearly impossible to reproduce (specialized hardware, intense computation requirements, sensitive data, human study, or other unavoidable reasons)	14
Difficult to reproduce because of unavoidable inherent complexity (e.g., requiring 300 million Markov chain Monte Carlo steps on each dataset, or needing months to do runs)	14
Reproducible with substantial tedious effort (e.g., individual download of a large number of datasets, hand coding of data into a new format, i.e., from an image, many archiving steps required)	5
Reproducible with substantial intellectual effort (e.g., methods well defined but required some knowledge of jargon or understanding of the field; or down the rabbit hole references to past articles required to reproduce; etc.)	5
Could reproduce with fairly substantial skill and knowledge (e.g., required GPU programing abilities to run code that wasn't given; translating complex models into MATLAB code; pseudo code with functions not detailed described in text into code; missing scripts)	23
Reproducible after tweaking (e.g., missing parameters required fiddling to find, missing modified code lines, missing arguments required for differing architecture; missing minor method step)	5
Minor difficulty in reproducing (e.g., installing a specialized library, converting to a different computational system)	18
Straightforward to reproduce with minimal effort	14

Table 4. Classification of reproducibility effort (n = 22)

code access information to a roughly equivalent sample of articles from before the policy implementation. To do this, we used the same selection criteria as our previous sample from 2011–2012 (volumes 331–336) to create a new sample from 2009–2010 (Volumes 325–330). Following the same methods, we obtained 956 titles from 2009–2010, from which we randomly selected 300 articles. After eliminating articles by duplicate authors and articles with no computational components, we were left with 213 articles from 2009–2010 and (unchanged) 204 from 2011–2012. Note that a similar sample size emerged in both periods, suggesting a similar pervasiveness of computational analysis over time among publications in *Science*.

We evaluated the 2009–2010 sample of 214 articles without contacting the authors. The 2011–2012 articles were inspected again to ensure comparability of results. We examined articles and supplemental materials for references to code and/or data used; whether information on how to get the underlying data and/or code appeared in the acknowledgements; whether the underlying data and/or code were mentioned; and whether further details needed for reproducibility, such as input parameters, workflow information, or other documentation, were mentioned. We present the results in Table 5. There was a minor improvement in citations, as 25% in 2009–2010 and 29% 2011–2012 articles cited code and/or data in the references section or in the supplementary references. There was a marked improvement for giving data deposition locations in the acknowledgements section: 29% in 2009–2010, to 48% in 2011–2012. However, code locations were rarely mentioned in acknowledgements in either sample: 4–5%. These results suggest progress in standards for data sharing and citation, and room for improvement in software sharing standards.

We evaluated whether the data and code could be obtained from the information provided in the article and supplemental materials section, and did not attempt replication this time. Data availability improved from 52% to 75% over the time period (Table 6). The improvement was less marked for providing code and software, however: 43% in 2009–2010 and 54% in 2011–2012. There was an improvement in the number of articles that shared both data and code. It is interesting that there were eight papers in our 2009–2010 sample without supplementary materials, while there was only one paper without supplementary materials in the 2011–2012 sample.

Methods

We selected all *Science* magazine publications after February 11, 2011, through June 30, 2012, to obtain a starting sample of 1,082 publications. We then removed from consideration 377 commentary, news, policy, data exhibit, and articles with duplicate authors. We randomly selected 300 papers of the remaining 705 and eliminated 96 noncomputational articles (e.g., theoretical results, experimental results), leaving 204 papers in the sample.

Survey Methods. This section describes our methods of requesting code and data for the 204 published articles in our sample and how we subsequently identified 56 potentially replicable journal articles. By inspection, we found that 24 of the 204 papers appeared to have made the complete set of code and data necessary for replication available via information included in the article and supplemental materials.

For the remaining 180 articles, we contacted the corresponding author via email, under Columbia University no. IRB-AAAK3050, which waived informed consent requirements. We used an Institutional Review Board (IRB)-approved template email to request the data and code from the corresponding author, customized to create a request for the specific data and code used to obtain the results in the paper but not provided in the article references or supplementary materials. The goal was to make a credible request from a researcher the author did not know and that was not easily dismissed as uninformed or lacking in seriousness. To avoid author name recognition, a Columbia student then sent the 180 authors the customized emails on April 26, 2013, using an automated email program. We felt that if computational artifacts were made available to a student, it seems reasonable they would be made available to other community members as well. This is an assumption worthy of further study, since requests from others, such as known colleagues, journal editors, or the general public, could garner a different level of response. The template email before article-specific customization of the data and code request is given in the associated GitHub repository listed below.

If we did not receive an answer after 2 wk, we sent a follow-up email request. We then permitted a second 2-wk interval to pass, and if we received no reply, we classified this article as not making available data and code. We received four late responses that attempted to supply data and code that were not included in our analysis because of our cutoff. The reason for this cutoff was twofold. We felt some time limit should apply, and 1 mo during the academic year seemed reasonable, and secondly, throughout the study, we sought to minimize the burden we placed on the researchers who were the subjects in our study. Therefore, we minimized our engagement with the researchers as much as reasonably possible. As noted earlier, it is possible that greater engagement with the author and a longer time horizon would procure more code and data.

Except for one case where we deemed the request well-founded, we did not respond to requests for further information, since we interpreted the *Science* policy narrowly as intending to make data and code available upon request only, and, as mentioned, we wished to minimize the time burden we placed on authors, and we wanted to keep the interaction with authors as uniform as possible across the study.

We received 131 timely responses to our 180 email requests; of those, 65 provided some data and/or code. We then evaluated these 89

Table 5. Changes in disclosure practices

Disclosure practice	2009–2010, %	2011–2012, %
Citations to data and/or code in references	25	29
Data location given in acknowledgements	29	48
Code location given in acknowledgements	4	5

Table 6. Materials availability via inspection

Materials availability	2009–2010, %	2011–2012, %
Most or all relevant data locations given	52	75
Most or all relevant software locations given	43	54
Some data and software locations given	25	45
All major software and data locations given	15	25
Code, scripts, parameters, documentation	10	12
No supporting materials available	4	1

"research compendia" (65 and the 24 articles which had provided access to data and code in the publication) and judged 56 papers to be potentially computationally reproducible by us. We use the term research compendia to refer to the bundle of these three digital scholarly outputs: publication and the associated data and code used to generate the results (2).

A flowchart representation of these steps is included in the GitHub repository associated with this publication.

Replication and Evaluation Methods. For the 56 articles deemed likely to be reproducible, we chose a random sample of 22 and attempted replication using the data and code the author provided, along with information in the article. The reproduction procedure started by recording all relevant figures, numerical and analytical conclusions, and the figure captions. We then attempted to deduce the computational methodology that was used for the collection and analysis of the available data to reproduce the figures and conclusions.

For the most part, useful replication methodology was rarely found in the article itself, with the exception of figure captions. Most details on methodology were found in the supplemental materials for each paper. Some articles cited analysis methods and models from previous publications. In these cases, we decided that we would not go more than two articles deep to find relevant parameters, data, or equations.

For each article, we filtered out the experimental details and extracted details on the computation and data analysis. All necessary data sources, software, and codes were listed, sorted, and downloaded for all articles deemed likely to be reproducible. The approximate time to accomplish this process for each paper was recorded, along with obstacles, licensing information, and the size of the collected data and code. Where data were too large to collect, it was noted. The data and code collection time varied according to the difficulty of the process.

For these 22 articles, all software was installed, and all data were examined (even for larger datasets). We noted where documentation on installing and running available codes was missing from both the paper and the location of the software, and continued with our best attempts at reproduction. We noted the cases where scripts and parameter files were missing. If the missing files could be recreated with some reasonable amount of effort ($<$100 lines of code) based on the information provided in the compendia, this was attempted.

For those processes which could be run within a reasonable amount of time ($<$1 wk) and with a reasonable amount of computational resources ($<$24 processors), codes were run with the input data and parameters used in the relevant articles. If the only bottleneck to reproducibility was our lack of sufficient computer resources, we did not count this against the article. In the cases where parameters were not provided, reasonable best guesses were attempted. For each reproduction attempt, code-run times, analysis time, and extra effort time (such as writing our own scripts or searching other papers due to missing documentation) were recorded. If the output of one impossible or impractical step was necessary as input for the next stage, sample input was used. This allowed us to verify the methodology, even where we might not have been able to verify the precise results.

It may not be the case that conclusions regarding the authors publishing in *Science* generalize to other communities, as norms and local expectations may differ. We also did not avail ourselves of the opportunity a reader has of alerting the journal editor when artifact requests go unfulfilled. We also note that it is likely some measure of selection bias exists in our study. *Science* is a multidisciplinary journal, and our findings may not generalize evenly to disciplinary journals: Some communities with more established sharing practices may expect higher percentages of sharing and reproducibility, and the converse may hold for communities just beginning their conversations. Another source of selection bias occurs in that authors who are more confident that their artifacts will replicate their results may be more likely to share when asked.

Under IRB requirements, we can only release aggregated data due to the potential for reidentification of study subjects. This means we are unable to make the raw data or code used in the replications publicly available. We have created a repository at https://github.com/ReproducibilityInPublishing/Science-2018 containing supplemental material, including the code and email templates that were used in the survey.

Conclusion

We were able to obtain data and code from the authors of 89 articles in our sample of 204, giving an estimate for the artifact recovery rate of 44% for articles published in *Science* shortly after the policy change: $(65 + 24/204)$ with a 95% bootstrap confidence interval of $[0.36, 0.50]$. Of the 56 articles that were then deemed potentially reproducible, we randomly chose 22 to attempt replication, and all but 1 of the 22 provided enough information that we were able to reproduce their computational findings (given sufficient resources and a willingness write some code). We estimate 95% $(21/22)$ of the articles deemed reproducible by inspection are computationally reproducible, so for the full sample, we estimate 26% will computationally reproduce $((56 * (1 - 1/22)))$ with a 95% bootstrap confidence interval for the proportion $[0.20, 0.32]$. We note limitations on our ability to draw broader conclusions regarding the potential drivers of reproducibility: Are some disciplines more likely to produce reproducible research? Do particular author characteristics imply greater reproducibility? A sample size of 21 reproduced articles limited our ability to carry out meaningful statistical inference across such a large set of possible drivers. A more direct comparison of disciplinary practices and other drivers of replication success is left to future work.

The comparison of artifact referencing and availability in Tables 5 and 6 lends itself to a simple difference in difference model, using a second similar reference journal as a control. This extension is also left as a future exercise, where the sample selection procedure described herein could be followed to generate a sample from a second journal with no policy change, and outcomes compared for before and after the 2011 policy change.

We found several serious shortcomings in usability and persistence for the digital artifacts associated with the publications in this study, suggesting that communities continue the conversation toward consensus on standards for documentation and metadata for data, code, and workflows that support findings in the scholarly record. The results from our survey show meaningful progress in standards for data sharing and citation, but much room for improvement for software citation standards, suggesting especially the need for improved community standards around the use and reuse of software.

Due to the gaps in compliance and the apparent author confusion regarding the policy, we conclude that, although it is a step in the right direction, this policy is insufficient to fully achieve the goal of computational reproducibility. Instead, we recommend that the journal verify deposit of relevant artifacts as a condition of publication (see, e.g., ref. 18). This is in compliance with Transparency and Openness Promotion Guidelines at Level 2 (6) and Recommendation 6 of the Reproducibility Enhancement Principles in ref. 4.

We recognize that some artifacts cannot be made publicly available for legal and other reasons, such as human subject research data, and exceptions can be disclosed in the publication (19, 20). There is progress on enabling greater sharing of sensitive data that promises to change this picture in the future. To address sensitive data, the notion of "quasireproducibility" was recently introduced to denote the availability of analysis code, along with simulated data that retains the key characteristics of the original data (21). Data perturbation techniques also present ways to protect confidential data and render it disclosable, while retaining its utility for scientific discovery and verification (22). Advances such as differential privacy enable queries on confidential data (23). New tools enabling automated data provenance capture and sharing are also reducing the effort to share (24, 25). Moving toward deposit of artifacts, at the time of publication, in open and trusted repositories appears to be the natural next journal policy step.

ACKNOWLEDGMENTS. We thank anonymous reviewers for constructive comments and criticism that improved the manuscript. This research was supported by Alfred P. Sloan Foundation Award PG004545 "Facilitating Transparency in Scientific Publishing" and NSF Award 1153384 "EAGER: Policy Design for Reproducibility and Data Sharing in Computational Science." See www.nsf.gov/awardsearch/showAward?AWD_ID=1153384. Since 2016, V.S. has been a member of the Transparency and Openness Promotion (TOP) Guidelines Coordinating Committee.

1. King G (1995) Replication, replication. *PS Polit Sci Polit* 28:444–452.
2. Gentleman R, Lang DT (2007) Statistical analyses and reproducible research. *J Comput Graphical Stat* 16:1–23.
3. Donoho DL, Maleki A, Rahman IU, Shahram M, Stodden V (2009) Reproducible research in computational harmonic analysis. *Comput Sci Eng* 11:8–18.
4. Stodden V, et al. (2016) Enhancing reproducibility for computational methods. *Science* 354:1240–1241.
5. National Academies of Sciences Engineering and Medicine (2017) *Fostering Integrity in Research*. (The National Academies Pres, Washington, DC).
6. Nosek BA, et al. (2015) Promoting an open research culture. *Science* 348:1422–1425.
7. Stodden V, Guo P, Ma Z (2013) Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLoS One* 8:e67111.
8. Hanson B, Sugden A, Alberts B (2011) Making data maximally available. *Science* 331:649.
9. American Association for the Advancement of Science (2011) Science journals: Editorial Policies. Available at www.sciencemag.org/authors/science-journals-editorial-policies. Accessed September 11, 2011.
10. Renear AH, Sacchi S, Wickett KM (2010) Definitions of *dataset* in the scientific and technical literature. *Proc Am Soc Inf Sci Technol* 47:1–4.
11. Stodden V, Borwein JM, Bailey DH (2013) "Setting the default to reproducible' in Computational Science Research. *SIAM News*. Available at https://sinews.siam.org/Aboutthe-Author/setting-the-default-to-reproducible-in-computational-science-research. Accessed February 16, 2018.
12. Peng RD (2011) Reproducible research in computational science. *Science* 334:1226–1227.
13. Stodden V, Bailey DH, Borwein J (2013) Set the default to 'open'. Notices of the AMS.
14. Stodden V (2009) The legal framework for reproducible scientific research: Licensing and copyright. *Comput Sci Eng* 11:35–40.
15. Ioannidis JPA, et al. (2008) Repeatability of published microarray gene expression analyses. *Nat Genet* 41:149–155.
16. Gilbert KJ, et al. (2012) Recommendations for utilizing and reporting population genetic analyses: The reproducibility of genetic clustering using the program structure. *Mol Ecol* 21:4925–4930.
17. Chang AC, Phillip L (2015) Is economics research replicable? Sixty published papers from thirteen journals say "usually not". *Finance Econ Discussion Ser* 83:1–26.
18. Fuentes M (2016) Reproducible research in JASA. AMSTAT News. Available at https://magazine.amstat.org/blog/2016/07/01/jasa-reproducible16/. Accessed February 16, 2018.
19. Lane J, Stodden V, Bender S, Nissenbaum H (2014) Enabling reproducibility in big data research: Balancing confidentiality and scientific transparency. *Privacy, Big Data, and the Public Good* (Cambridge Univ Press, New York), pp 112–132.
20. Wolf LE, et al. (2015) Certificates of confidentiality: Protecting human subject research data in law and practice. *J L Med Ethics* 43:594–609.
21. Coughlin S (2017) Reproducing epidemiologic research and ensuring transparency. *Am J Epidemiol* 186:393–394.
22. Muralidhar K, Sarathy R (2012) Perturbation methods for protecting numerical data: Evolution and evaluation. *Handbook of Statistics* 28:513–531.
23. Lane J, Stodden V, Bender S, Nissenbaum H (2014) Differential privacy: A cryptographic approach to private data analysis. *Privacy, Big Data, and the Public Good* (Cambridge Univ Press, New York), pp 296–320.
24. Carata L, et al. (2014) A primer on provenance. *ACM Queue* 12.
25. Brinckman A, et al. (February 10, 2018) Computing Environments for Reproducibility: Capturing the "Whole Tale." *Future Gener Comp Sy*, 10.1016/j.future.2017.12.029.