

Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology



Pepijn Obels¹, Daniël Lakens¹, Nicholas A. Coles²,
Jaroslav Gottfried³, and Seth A. Green⁴

¹Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology; ²Department of Psychology, University of Tennessee; ³Department of Psychology, Masaryk University; and ⁴Department of Psychology, Princeton University

Advances in Methods and
Practices in Psychological Science
2020, Vol. 3(2) 229–237
© The Author(s) 2020



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245920918872
www.psychologicalscience.org/AMPPS



Abstract

Ongoing technological developments have made it easier than ever before for scientists to share their data, materials, and analysis code. Sharing data and analysis code makes it easier for other researchers to reuse or check published research. However, these benefits will emerge only if researchers can reproduce the analyses reported in published articles and if data are annotated well enough so that it is clear what all variable and value labels mean. Because most researchers are not trained in computational reproducibility, it is important to evaluate current practices to identify those that can be improved. We examined data and code sharing for Registered Reports published in the psychological literature from 2014 to 2018 and attempted to independently computationally reproduce the main results in each article. Of the 62 articles that met our inclusion criteria, 41 had data available, and 37 had analysis scripts available. Both data and code for 36 of the articles were shared. We could run the scripts for 31 analyses, and we reproduced the main results for 21 articles. Although the percentage of articles for which both data and code were shared (36 out of 62, or 58%) and the percentage of articles for which main results could be computationally reproduced (21 out of 36, or 58%) were relatively high compared with the percentages found in other studies, there is clear room for improvement. We provide practical recommendations based on our observations and cite examples of good research practices in the studies whose main results we reproduced.

Keywords

reproducibility, Registered Reports, data sharing, open science, open data, open materials

Received 9/11/19; Revision accepted 2/27/20

Researchers are currently exploring ways to make science more open and transparent. Among the novel developments that are part of this effort are preregistration, preprints, and open peer review. In addition, an increasing number of journals, funders, and researchers are beginning to expect that data, materials, and analysis code will be shared by default with scientific publications (e.g., Morey et al., 2016). Sharing data and analysis code with scientific publications allows other people to more easily reproduce, check, and build on existing work. This requires the development of new skills and best practices because most scientists have not received training in how to make their work reproducible. It is important to evaluate how data and code

are currently being shared, and how easy it is to reproduce analyses reported in the published literature, to learn what can be improved. With this goal in mind, we attempted to computationally reproduce the main results of Registered Reports published in the psychology literature.

It is desirable that research is reproducible. Data availability has the potential to make science more efficient by facilitating the reuse of data. The availability of analysis code makes it possible for peers to check and correct

Corresponding Author:

Daniël Lakens, ATLAS 9.402, 5600 MB, Eindhoven, The Netherlands
E-mail: D.Lakens@tue.nl

published findings. According to Kitzes, Turek, and Deniz (2017), computational reproducibility means that “a second investigator (including the original researcher in the future) can recreate the final reported results of the project, including key quantitative findings, tables, and figures, given only a set of files and written instructions” (p. xxii). For scientific research to be computationally reproducible, the data and code need to be shared.

However, the availability of data and code in itself is not enough. Articles need to provide links to these materials so that readers know where to find them. Preferably, the data should be available in a format that can be read by open-source software. Variables must be described and labeled (e.g., in a codebook), and code should be annotated. Finally, the results reported in the article should be reproducible; that is, it should be possible to compute these results using the available data and code.

Recently, scholars have started to empirically examine the extent to which data for published articles are shared, and when they are, whether it is possible to reproduce the data analyses reported. Hardwicke et al. (2018) examined the analytic reproducibility of 35 articles published in the journal *Cognition*. Target outcomes that supported the identified substantive findings of 11 articles could be reproduced by independently writing analysis code, without assistance from the original authors, and in each of 13 sets of target outcomes, at least one outcome could not be reproduced even with the original authors' assistance. Hardwicke et al. estimated that it took between 2 and 25 hr per article to complete the reproducibility checks, but they did not record the exact time. Stockemer, Koehler, and Lentz (2018) analyzed reproducibility in all articles published in 2015 in three political-science journals. They e-mailed authors for the code and data, which they received for 71 articles. The results of 1 article could not be reproduced because of a lack of a software license, and output for 16 articles' findings could not be obtained even with access to the required software. Thirty-two sets of results could be exactly reproduced, and 19 could be reproduced with slight differences; the results for the remaining 3 articles were significantly different from the original results. Stodden, Seiler, and Ma (2018) analyzed data availability for articles in the journal *Science* in 2011 and 2012 and found that 26 of 204 articles (or 13%) provided information to retrieve data, code, or both without contacting the authors. Stodden et al. e-mailed authors for data and code and estimated that results for 26% of the data sets they had acquired were computationally reproducible. These studies reveal that there is clear room for improvement in how reproducible published results are.

We set out to examine the data availability and reproducibility for Registered Reports published in psychological science. Our main interest was to examine the computational reproducibility of the main analyses reported in published articles, without contacting the original authors. One of the main benefits of sharing data and code alongside an article (compared with making these files available upon request) is that results can be reproduced and data reused even if the original author can no longer be reached.

Registered Reports are a novel development in psychology. Before data collection commences, the introduction and methods are peer-reviewed, after which authors can receive an in-principle acceptance. This means that the article will be published as long as the authors follow their preregistered data-collection and -analysis plan (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014; Nosek & Lakens, 2014). The population of Registered Reports in psychology is still relatively small (118 Registered Reports had been published as of June 5, 2018), so it is possible to examine this population in full.

The novelty of Registered Reports may attract early adopters who are also exploring other novel developments aimed at improving research practices in psychology, such as data and code sharing. In addition, one journal that publishes Registered Reports (*Royal Society Open Science*) requires authors to deposit data and code, and several other journals that publish Registered Reports strongly encourage data or code sharing. We expected researchers who publish Registered Reports to be likely to share data and code in public repositories and to embrace computational reproducibility. To evaluate the reproducibility of findings published in Registered Reports, we examined if data could be located, were available at the indicated location, could be opened in open-source or accessible software, were documented well enough to be understandable, and could be used to reproduce the main analyses reported in the published article.

Our main objective was to examine how reusable data and code underlying Registered Reports are, given solely the information provided in those articles, so that we could identify how the reproducibility of results reported in these articles could be improved. We examined how many authors shared data and code without our solicitation and the extent to which we could reproduce reported analyses without contacting the original authors. While attempting to reproduce the results reported in Registered Reports, we kept track of factors that facilitated reproducibility or that made reproducing results more difficult. We report these qualitative findings with an aim to highlight how current practices can be improved.

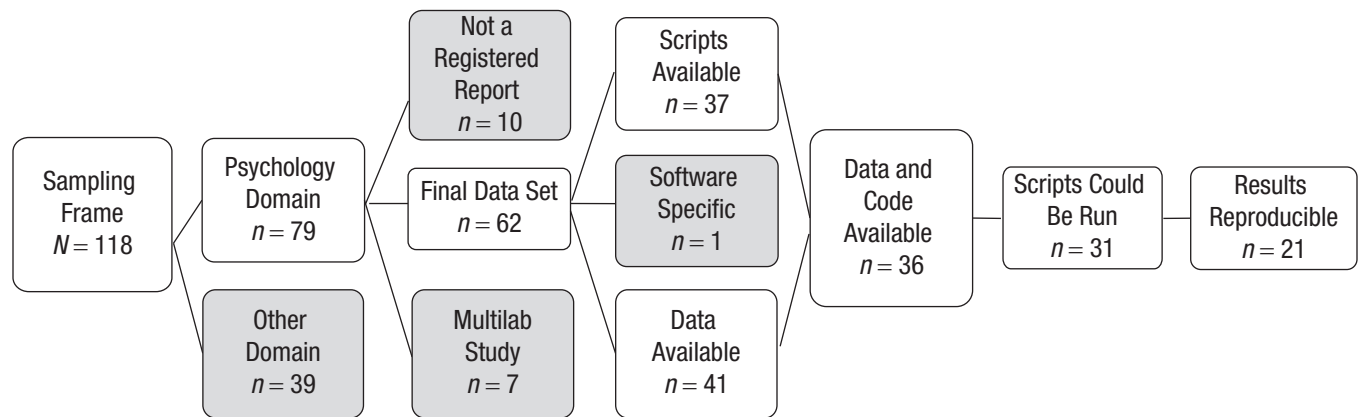


Fig. 1. Flowchart showing the number of Registered Reports that met each of the criteria in this study.

Disclosures

All data and code for this article are provided in an Open Science Framework (OSF) repository, at <https://osf.io/suqz3/>. We report all measured variables and all analyses we conducted.

Method

To find Registered Reports published in the psychology literature, we drew from a database of Registered Reports maintained by the Center for Open Science.¹ At the start of this project (June 5, 2018), this database consisted of 118 published Registered Reports from 2014 to 2018. Seventy-nine of these were published in the psychology literature (see the flowchart in Fig. 1 for the number of articles that met each of the criteria in this study). We limited our analysis to studies performed by single groups, because such articles are most representative of the researchers' current work, and excluded 7 large-scale collaborations in which dedicated team members were responsible for making analyses reproducible, such as Registered Replication Reports (Hagger et al., 2016) and Many Labs projects (R. A. Klein et al., 2014; R. A. Klein et al., 2018). Upon further inspection of the 72 articles left in the data set, we found that 10 were not formally Registered Reports. This left 62 articles in our sample. When evaluating whether we could reproduce the original results, we limited ourselves to statistical software packages that we had experience with: R (R Core Team, 2017), SPSS, Python (Python Software Foundation, <http://www.python.org>), MATLAB (The MathWorks, Natick, MA), and JASP (jasp-stats.org); 1 additional study was excluded because it required expertise in software packages we were not trained in (e.g., dedicated software for electroencephalography, EEG).

Results

We set out to reproduce the findings of the 62 Registered Reports that met our inclusion criteria. For each article, we coded whether the data and code were (a) linked to the article (or could be found by searching OSF), (b) available, (c) not software-specific, and (d) understandable and whether (e) the reported results were reproducible. These five categories were inspired by FAIR data principles concerning the *findability*, *accessibility*, *interoperability*, and *reusability* of data and code, but we did not explicitly code whether articles adhered to the exact definitions of the four FAIR principles (Wilkinson et al., 2016). Our main aim was to examine the reproducibility of results, and adhering to FAIR criteria requires meeting more stringent requirements, for example, concerning the presence of metadata (which were missing for all data sets in our sample). We considered data or code to be linked to the article when it included a unique link to the data or scripts. Ideally, such a link consists of a stable digital object identifier (DOI). A hyperlink to a website that contains the data and script also suffices, although hyperlinks are known to break over time (Gertler & Bullock, 2017).

Of the 62 articles, 45 provided a link to the relevant data, code, or both. Linking to data or code does not mean that the data or code is actually available, and the absence of a link does not mean that the data or code is unavailable. The link for 1 article no longer worked (which highlights the benefit of using a stable DOI). For 3 articles, the link still worked, but there were no data at the linked destination. In another 3 cases, there was no link to the data in the article, but we were nevertheless able to find the data on OSF when we searched for the title of the article. For 1 report, the linked data were not available because they were embargoed until a future date.

The frequency of sharing was in line with our prediction that authors of Registered Reports would be relatively likely to also adopt other open-science practices; the authors of 43 of the 62 articles in our final data set shared at least some of their data and code (69.4%). In comparison, after the journal *Cognition* introduced a mandatory data-sharing policy, 136 out of 174 articles (78.2%) had a data-availability statement, 85 out of 174 articles (48.9%) had reusable data, and only 18 out of 174 articles (10.3%) provided the analysis code (Hardwicke et al., 2018). For 37 articles in our final data set, the available code contained the statistical analyses, and for 41, all the data files required to reproduce the reported results were available (for 36 articles, both data and code were available).

We also coded the extent to which data and code were specific to software that was not freely available. When open-source software is used, the analyses can be reproduced by anyone with time, a computer, and access to the Internet. When proprietary software is used, results might still be reproducible in principle, but could require more effort to do so. For example, SPSS produces proprietary .sav and .sps files. However, .sav files can be opened in R, and .sps files can be opened by a text editor, and the code can be rewritten as long as it is annotated well enough to be recoded in R. When examining whether analyses were reproducible, we used only the same software packages that had been used by the original authors. The data files of 1 article, on an EEG study, consisted of .eeg, .vhdr, and .vmrk files, which require dedicated EEG software and could not be reproduced (we also could not find the analysis code for this article).

One of the reasons to share data is to allow other researchers to reproduce the reported results. Another important reason is to enable other researchers to reuse the data. If the data can be understood by others, they can be used to answer novel research questions. This is one of the reasons why it is considered best practice to describe a data set's variables in a codebook. If the variables are not clearly described (e.g., they are identified by abbreviations that make sense only to the original researcher), other researchers will not be able to reuse the data to answer novel questions. In our analyses, data were scored as "understandable" when all variables were clearly named (e.g., "Condition") and the values for variables were labeled (e.g., 0 = control, 1 = experimental). Out of the 44 data sets that were in a format that was not software-specific, only 24 were described in enough detail to be understandable. This highlights the importance of adding a codebook with a data file.²

Finally, we examined how many of the 36 articles with data and code available had results that could be

reproduced. It is possible that running the code on the data reproduces all analyses, even when the data file itself is not understandable (i.e., the data columns are not labeled). Two authors per article coded the SPSS, R, MATLAB, Python, and JASP analyses for (a) the executability of the script and (b) the reproducibility of the results. After the initial coding, interrater reliability was low (60% agreement on executability and 55% agreement on reproducibility for SPSS scripts, 75% agreement on executability and 56% agreement on reproducibility for R scripts). This initial low agreement provided two important insights about the definition of reproducibility and executability, on the one hand, and the role of expertise, on the other hand.

When coding whether the script could be executed and the results could be reproduced, we used a dichotomous classification ("yes" or "no"), but the coders often reported "partial" reproducibility. Code often needed minor adjustments to run on the data, such as changing file locations or loading packages in R, and the coders sometimes took different approaches to how much they adjusted the code to make it run on the data (for detailed comments, see our data file at our OSF project page). When judging if the code ran on the data, we allowed for minor errors, but categorized code as not running when it was unclear how analysis code related to data files or when there were a substantial number of errors even after attempts to make minor corrections. Furthermore, we did not preregister this study and had no clearly defined coding scheme based on pilot data. As a consequence, the coders initially used different thresholds of reproducibility—for example, whether every single result or only the main results reported could be reproduced using the code and data. After evaluating our initial coding round, we decided to consider an article's results reproducible when we could get the same main results as represented in the article with at most minor changes to the analysis scripts. This means that we considered the analysis reproducible even if, in the absence of a codebook, the coders had to search through the analysis code and data set to identify how variable and label names related to the results reported. Furthermore, we changed folder locations when needed and installed and loaded required libraries. Finally, even though some figures were generated in R and contained relevant information (e.g., the pattern of means), we did not require all figures to be reproducible. For each study reported in an article, we identified main results (i.e., any reported descriptive statistics and statistical tests) on the basis of the research question highlighted in the title and abstract. Approximately half of the articles reported replication studies; in these cases, the main analysis was explicitly stated and based on a previous study. In

the remaining articles, the main research question was often stated clearly in a “confirmatory analyses” section. Nevertheless, some arbitrary judgments were required when the coders decided which analyses were part of the main results.

For each article, two coders examined whether the code could be run on the data; a third coder attempted to reproduce the results of five articles for which some uncertainty remained. There were differences between the coders in how much expertise they had with R, SPSS, and JASP (P. Obels had less experience, and J. Gottfried, N. A. Coles, and D. Lakens had more experience; S. A. Green was responsible for all MATLAB and Python code). The more expertise the coders had, the easier it was for them to reproduce findings (which lowered interrater reliability). This raises the question of what level of required skill should be the threshold beyond which results reported in a scientific article should not be considered reproducible. The required experience might be difficult to quantify. Our results concerning the reproducibility of results are based on which results could be reproduced by a Ph.D. student who had experience with the statistical software and had been educated in the same scientific discipline, and we feel this is a reasonable standard. We considered a study’s results to be reproducible as long as the more experienced coder could reproduce the main results. All the authors collectively discussed disagreements, which on several occasions led to clarifying ambiguities in rating strategies and correcting mistakes in the analytic process (e.g., overlooked script files). The final ratings presented here have been discussed and approved by all the authors, but are not completely flawless, and we expect that different teams of coders would reach slightly different conclusions (i.e., perfect reliability would be extremely difficult to achieve). It may bear repeating that the main goal of our analysis was to examine how reusable data and code underlying Registered Reports are, so as to evaluate where there is room for improvement and to identify practices that researchers can use to make their work more reproducible.

Of the original 62 articles, 36 had the analysis code and data available. We were able to run the code for 31 out of the 37 articles that had the code available. R was used as a coding language for 13 of the articles, and the scripts for 10 of them could be run on the data; SPSS was used for 17 of the articles, and the scripts for 15 of them could be run; both SPSS and R were used for 3 of the articles, and all these scripts could be run. JASP does not separate the data and code, but instead stores both in a single JASP file. This means that the analyses are always directly linked to the code for every output. It is always clear which settings were used to generate results. Because of this useful feature, we were

able to reproduce the analyses of the 2 articles that relied on JASP. Finally, the code for 1 article using MATLAB could be run on the data. Being able to run the code on the data does not imply that all the main analyses reported in an article can be correctly reproduced. Some analyses reported might not be part of the code or the output. We found that for 21 out of the 36 articles (58.33%) for which data and code were available, these could be used to reproduce the main results reported in the article.

For the 15 articles whose results could not be reproduced, the main reasons were as follows: (a) In 8 cases, code to reproduce some values (e.g., dedicated code to run a macro in SPSS) was missing; (b) in 6 cases, the code gave errors (e.g., variables in the data set were missing, or functions did not run as expected), and (c) in 1 case, the results might have been reproducible, but the code was so complex that after 40 min only a small part of the results could be reproduced, and it was judged that the time needed to check the rest of the results fell outside of a reasonable time span. In the articles whose results were coded as reproducible, two small errors were observed (one value of 0.89 was rounded to 0.88, and a value that should have been 0.03 was reported as 0.06, probably because of a typo), but given that dozens of other values in these articles were reproduced perfectly, these errors were not considered severe enough to deem the main results not computationally reproducible.

After the data and code had been downloaded, and the reported results had been identified by looking through the article, we recorded the time it took to reproduce analyses. The average time to reproduce analyses in R was 27.08 min ($SD = 28.55$) for the first coder and 32.50 min ($SD = 20.95$) for the second coder. Reproducing the SPSS analyses took on average 17.35 min ($SD = 9.54$) for the first coder and 25.50 min ($SD = 9.72$) for the second coder. Most of the time was spent matching output from the statistical analyses to the analyses reported in the article. This suggests that even if results are reproducible, the organization of the output and the relation of the output to the published article, can often be improved.

Discussion

We analyzed 62 Registered Reports to examine how many authors shared their data and code and how often the main results reported could be reproduced. In total, 36 out of the 62 articles (or 58.06%) shared the underlying data and the code that was used to generate the results. Authors of Registered Reports in psychology seem to share data and code relatively often compared with authors of articles in political science (17.93%;

Stockemer et al., 2018). Compared with other types of articles in psychology, Registered Reports have relatively high rates of data and code sharing, as well as reproducibility (taking into account that we reproduced results without contacting the original authors). The reproducibility rate we found was higher than the 31% rate observed by Hardwicke et al. (2018), and both data sharing and reproducibility were higher than in the sample of articles from the journal *Science* in 2011 and 2012 analyzed in Stodden et al. (2018).

Nevertheless, our results indicate that there is clear room for improvements in the computational reproducibility of Registered Reports. One of the main goals of our project was to identify ways to improve the reproducibility of published articles. We encountered several common issues that made results reported in Registered Reports difficult to reproduce (cf. Hardwicke et al., 2018). On the basis of our observations, we recommend that researchers in psychology should focus on four areas to improve reproducibility, namely, (a) adding a codebook to data files, (b) annotating code so that it is clear what the code does and clearly structuring code (e.g., using a README file) so that others know which output analysis code creates, (c) checking whether the code shared still reproduces all analyses after revisions during the peer-review process, and (d) listing packages that are required in R and the versions used at the top of the R file. We discuss each of these points in turn and cite examples of good practices that we encountered.

First, data are easier to understand and more reusable if variables and their values are clearly described, for example, in a codebook. Researchers should ensure that their codebook and variable names are in the same language as the article. Furthermore, when there are multiple data files, researchers should provide a clear description of what each data file contains, for example, in a README file in the root directory of the data folder. (Le, 2018) has provided useful guidelines to create codebooks in his *Open Science Manual*. A good example of a codebook can be found as part of the materials of Wesselmann et al. (2014). Creating a codebook should be considered a best practice for sharing data.

Second, code should be well annotated, so that it is understandable for researchers who did not write the code. Good annotation makes clear what the analysis code does, in which order scripts should be run if there are multiple scripts (e.g., to preprocess the raw data, compute sum scores, analyze the results, and generate graphs), and which output each section of the analysis code generates. A good example of well-annotated code can be found in the materials of Weston and Jackson (2018). Annotation helps to make clear how the analysis code relates to the analyses reported in the

article, to make it easier for other researchers to identify which code generates which results. For one article that we coded as not reproducible, there was too much unstructured code, and analyses took too long to run, so that we decided that the results were not reproducible with a reasonable amount of effort. Explicitly linking code in the analysis script to the final article also helps researchers to check whether all results in the article are reproduced by the shared code. An example of a data-analysis file that clearly links the code to the final article can be found in the materials of Voorspoels, Bartlema, and Vanpaemel (2014). If some of the code is for analyses not included in the article (e.g., assumption checks, exploratory analyses), this should be stated explicitly. The structure of analysis scripts can often be improved by creating different sections in the code, or creating different files for different parts of the data analysis (e.g., data cleaning, data preparation, exploratory data analysis, and confirmatory data analysis).

Third, we recommend that researchers perform a final check after peer review has been completed to make sure that any changes in the code introduced during the peer-review process are reflected in the shared data and code.

Finally, on the basis of our experiences, we have several specific recommendations for data analyzed in R. First, most code in R relies on specific libraries (also called packages). All the packages that the code needs to run should be listed at the top of the script. Because packages are updated, it is necessary to report the version numbers of packages that were used (e.g., by using *packrat*; Ushey, McPherson, Cheng, Atkins, & Allaire, 2018) or copying the output of the `sessionInfo()` function as a comment in the script). Folder names and folder structures differ between computers, and therefore it is important to use relative locations (and not, e.g., “c:/user/myfolder/code”). RStudio (rstudio.com) and the *here* package (Müller, 2017) provide an easy way to use relative paths. When multiple scripts are used in the analysis, a README file should indicate the order in which the scripts should be run. R Markdown (Allaire et al. 2020) files provide a useful way to share clearly annotated code and structure the different steps in the data analysis, for example, as done by Campbell et al. (2018).

When we tried to reproduce the results of SPSS scripts, the biggest issue was the often confusing and unclear structure of the scripts. Large portions of the scripts were not annotated, and it was unclear which results they should produce. Often, the descriptive, confirmatory, and exploratory analyses were not easily distinguishable because of an overall lack of structure. The absence of understandable variable and value labels in more than half of all the SPSS scripts hindered

our attempts to reproduce these results. Often, the only time-efficient way to check if an article's results were reproducible was to run the whole script and try to identify specific p values or effect sizes from the article in the SPSS output. SPSS users should take care to clearly organize their analysis scripts by adding comments or a README file that links results generated by the SPSS scripts to the analyses reported in the article. Another frequent problem was missing or incorrectly labeled variables in the data set, so the scripts could not run properly. We suspect that this was the result of authors updating or modifying either their data sets or their scripts during the publication process. Such discrepancies could be easily detected if a second author attempted to reproduce the analyses in the final report before data and scripts are shared publicly.

We limited our analysis to Registered Reports because we thought that this article format might be used by people who are early adopters of innovations in science and would therefore be particularly likely to also share data and code. We found that the rate at which data and code were shared in our sample was high relative to the rates observed in other reproducibility-focused reviews (e.g., Hardwicke et al., 2018; Stockemer et al., 2018), but we do not have data that give insights into the motivations of these authors. Registered Reports are written by a diverse set of researchers, working in different subfields in psychology, and it would be interesting for future research to qualitatively examine the motivations of Registered Report authors for sharing or not sharing their data and code. There are several good reasons why some data should not be shared, and in such cases, researchers should be encouraged to explain their reasons (Morey et al., 2016).

The main aim of this project was not to precisely estimate reproducibility rates, but rather to see what current standards are and how the reproducibility of results reported in research articles using the Registered Report format could be improved. Our sample size is small, and it is doubtful whether a precise estimate of the reproducibility of results in Registered Reports is of much value. Data and code sharing are relatively new, researchers typically lack training in reproducible data analysis, and therefore the main contribution of this article is the identification of common problems that can be remedied. We have provided some suggestions and examples of better practices that should make the results in published articles more reproducible.

In addition to the recommendations we have provided, novel technological solutions might improve the reproducibility of results reported in research articles. For example, Code Ocean is an online, cloud-based, computational reproducibility platform (Clyburne-Sherin, Fei, & Green, 2018). It provides a code environment (or

container) that runs online, which means that researchers using Code Ocean do not have to download data, code, or software, but can analyze the data in their browser. It is not currently possible to use SPSS within Code Ocean, but for R code, it solves the problem of package versions (because the container uses the versions specified by the researchers) and file locations.³ Other platforms in the reproducibility space include Whole Tale (Brinckman et al., 2019), "a research environment that captures and, at the time of publication, exposes salient details of the entire research process via access to persistent versions of the data and code used, provenance, and data lineage" (p. 855), and Binder (Project Jupyter et al., 2018), an open-source, browser-based tool for creating and sharing reproducible environments. Another useful technology is R Markdown, which enable researchers to write fully executable manuscripts. R Markdown files load the raw data and allow researchers to compute each number reported in the article from the data, instead of copying and pasting values. This means that, as long as the data and required packages can be loaded, all reported numbers can be reproduced. This saves time when researchers try to match the analysis code's output to reported results, and thus speeds up the process of checking whether all results reported in the article are reproduced. The accepted manuscript for this article is an example of a reproducible R Markdown file.⁴ Additional solutions that help researchers to share reproducible analyses may become available in the future.

Finally, journals that value reproducibility might find it worthwhile to check whether the data and analysis code shared with a submission can be used to reproduce the results. The average time it took our team to check that the analysis code could reproduce the reported results was 24 min. This is slightly shorter than the time it took Hardwicke et al. (2018), who estimated (without keeping track of the time explicitly) that checking reproducibility and preparing a reproducibility report took between 2 and 25 person-hours, depending on whether the article eventually fell in the reproducible or not-reproducible category, and whether an author's assistance was needed. One major difference between our approach and theirs is that we did not write our own code to analyze the data, as Hardwicke et al. did, but simply ran the code written by the original authors on the shared data. We also did not create a reproducibility report for each article. Documenting the process of reproducing reported results adds transparency and allows other researchers to check the decisions about every value. Whether such a level of detail is worth the additional time invested in documenting each reported value is a cost-benefit analysis that journals should undertake for themselves. The required

time might be reduced by explicitly asking authors to submit files in a format or structure that facilitates such checks, or by automating part of the work that is needed to check the reproducibility of results. Overall, we feel that the time required for a basic check of the computational reproducibility of articles (i.e., a check on whether the *main* results are reproduced by the analysis scripts, without documenting this process at the level of each individual number) is a surmountable hurdle for journals, and would substantially improve the computational reproducibility of the published literature.

The best route to progress, in addition to developing novel technologies, will probably be to develop standards within research communities and educate researchers about best practices that guarantee reproducibility (for recent examples, see O. Klein et al., 2018; Liu & Salganik, 2019). Most researchers are not trained in reproducible data analysis and cannot be expected to invent best practices from scratch. As good examples appear in the published literature over time, and best practices within subdisciplines crystallize, standards that emerge should improve reproducibility and allow researchers to share data and code in such a way that others with basic scientific training can reproduce their results and reuse their data.

Transparency

Action Editor: Alexa Tullett

Editor: Daniel J. Simons

Author Contributions

P. Obels and D. Lakens developed the idea for this project. All the authors contributed data by reproducing analyses. P. Obels drafted the initial version of the manuscript, D. Lakens wrote the final version, and all the authors revised the manuscript.

Declaration of Conflicting Interests

S. A. Green worked at Code Ocean during the writing of this manuscript and joined the project after its Code Ocean component was already submitted and published; he did not write any of the text concerning Code Ocean. The authors declared that there were no other potential conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported by the Netherlands Organisation for Scientific Research (NWO) VIDI Grant 452-17-013.

Open Practices

Open Data: <https://osf.io/kwr37>

Open Materials: <https://osf.io/kwr37>

Preregistration: not applicable


All data and materials have been made publicly available via the Open Science Framework and can be accessed at <https://osf.io/kwr37>. The complete Open Practices Disclosure for this article can be found at <http://journals.sagepub.com/doi/suppl/10.1177/2515245920918872>. This article

has received badges for Open Data and Open Materials. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>.



ORCID iDs

Daniël Lakens  <https://orcid.org/0000-0002-0247-239X>

Nicholas A. Coles  <https://orcid.org/0000-0001-8583-5610>

Notes

1. The database with published Registered Reports can be found at <https://www.zotero.org/groups/479248/osf/items/collectionKey/KEJP68G9>.
2. For an explanation of how to use the *codebook* R package to create machine-readable codebooks, see Arslan (2019). Note that we used this package to create the codebook that accompanies this article.
3. For a Code Ocean capsule reproducing the accepted manuscript for this article, see <https://doi.org/10.24433/CO.4275368.v1>.
4. This file is available at https://github.com/Lakens/reproducing_registered_reports/blob/master/manuscript_version_2/reproducing_registered_reports.Rmd.

References

- Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., . . . Iannone, R. (2020). rmarkdown: Dynamic documents for R (R package Version 2.1) [Computer software]. Retrieved from <https://CRAN.R-project.org/package=rmarkdown>
- Arslan, R. C. (2019). How to automatically document data with the *codebook* package to facilitate data reuse. *Advances in Methods and Practices in Psychological Science*, 2, 169–187. doi:10.1177/2515245919838783
- Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M. B., Kowalik, K., . . . Turner, K. (2019). Computing environments for reproducibility: Capturing the “whole tale.” *Future Generation Computer Systems*, 94, 854–867.
- Campbell, L., Balzarini, R. N., Kohut, T., Dobson, K., Hahn, C. M., Moroz, S. E., & Stanton, S. C. E. (2018). Self-esteem, relationship threat, and dependency regulation: Independent replication of Murray, Rose, Bellavia, Holmes, and Kusche (2002) Study 3. *Journal of Research in Personality*, 72, 5–9. doi:10.1016/j.jrp.2017.04.001
- Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. J. (2014). Instead of “playing the game” it is time to change the rules: Registered Reports at *AIMS Neuroscience* and beyond. *AIMS Neuroscience*, 1, 4–17. doi:10.3934/Neuroscience.2014.1.4
- Clyburne-Sherin, A., Fei, X., & Green, S. A. (2018). Computational reproducibility via containers in social psychology. *PsyArXiv*. doi:10.31234/osf.io/mf82t
- Gertler, A. L., & Bullock, J. G. (2017). Reference rot: An emerging threat to transparency in political science. *PS: Political Science & Politics*, 50, 166–171. doi:10.1017/S1049096516002353

- Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., . . . Zwiener, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*, 546–573. doi:10.1177/1745691616652873
- Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsson, G., Banks, G. C., Kidwell, M. C., . . . Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal *Cognition*. *Open Science, 5*(8), Article 180448. doi:10.1098/rsos.180448
- Kitzes, J., Turek, D., & Deniz, F. (2017). *The practice of reproducible research: Case studies and lessons from the data-intensive sciences*. Oakland: University of California Press.
- Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., . . . Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology, 4*, Article 20. doi:10.1525/collabra.158
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology, 45*, 142–152. doi:10.1027/1864-9335/a000178
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1*, 443–490. doi:10.1177/2515245918810225
- Le, B. (2018). *Open science manual*. Retrieved from <https://bit.ly/2w2F6Xu>
- Liu, D., & Salganik, M. (2019). Successes and struggles with computational reproducibility: Lessons from the Fragile Families Challenge. *SocArXiv*. doi:10.31235/osf.io/g3pdb
- Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., . . . Zwaan, R. A. (2016). The Peer Reviewers' Openness Initiative: Incentivizing open research practices through peer review. *Royal Society Open Science, 3*(1), Article 150547. doi:10.1098/rsos.150547
- Müller, K. (2017). here: A simpler way to find your files (R package Version 0.1) [Computer software]. Retrieved from <https://CRAN.R-project.org/package=here>
- Nosek, B. A., & Lakens, D. (2014). Registered Reports. *Social Psychology, 45*, 137–141. doi:10.1027/1864-9335/a000192
- Project Jupyter, Bussonnier, M., Forde, J., Freeman, J., Granger, B., Head, T., . . . Willing, C. (2018). Binder 2.0 - reproducible, interactive, sharable environments for science at scale. Retrieved from https://conference.scipy.org/proceedings/scipy2018/pdfs/project_jupyter.pdf
- R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Stockemer, D., Koehler, S., & Lentz, T. (2018). Data access, transparency, and replication: New insights from the political behavior literature. *PS: Political Science & Politics, 51*, 799–803. doi:10.1017/S1049096518000926
- Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences, USA, 115*, 2584–2589. doi:10.1073/pnas.1708290115
- Ushey, K., McPherson, J., Cheng, J., Atkins, A., & Allaire, J. J. (2018). packrat: A dependency management system for projects and their R package dependencies (R package Version 0.5.0) [Computer software]. Retrieved from <https://CRAN.R-project.org/package=packrat>
- Voorspoels, W., Bartlema, A., & Vanpaemel, W. (2014). Can race really be erased? A pre-registered replication study. *Frontiers in Psychology, 5*, Article 1035. doi:10.3389/fpsyg.2014.01035
- Wesselmann, E. D., Williams, K. D., Pryor, J. B., Eichler, F. A., Gill, D. M., & Hogue, J. D. (2014). Revisiting Schachter's research on rejection, deviance, and communication (1951). *Social Psychology, 45*, 164–169. doi:10.1027/1864-9335/a000180
- Weston, S. J., & Jackson, J. J. (2018). The role of vigilance in the relationship between neuroticism and health: A registered report. *Journal of Research in Personality, 73*, 27–34. doi:10.1016/j.jrp.2017.10.005
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, Article 160018. doi:10.1038/sdata.2016.18