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

Pepijn Obels[1], Daniel Lakens[1], Nicholas A. Coles[2], Jaroslav Gottfried[3], & Seth Ariel Green[4]

[1] Eindhoven University of Technology, The Netherlands
[2] University of Tennessee, Knoxville, USA
[3] Masaryk University, Brno, Czech Republic
[4] Code Ocean, New York, USA

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 re-use or check published research. However, these benefits will only emerge if researchers can reproduce the analysis reported in published articles and if data is annotated well enough so that it is clear what all variables mean. Because most researchers are not trained in computational reproducibility, it is important to evaluate current practices to identify practices that can be improved. We examined data and code sharing for Registered Reports published in the psychological literature between 2014 and 2018, and attempted to independently computationally reproduce the main results in each article. Of the main results from 62 articles that met our inclusion criteria, data were available for 41 articles, and analysis scripts for 37 articles. For the main results in 36 articles that shared both data and code we could run the scripts for 31 analyses, and reproduce the main results for 21 articles. Although the articles that shared both data and code (36 out of 62, or 58%) and articles for which main results could be computationally reproduced (21 out of 36, or 58%) was relatively high compared to other studies, there is clear room for improvement. We provide practical recommendations based on our observations and link to examples of good research practices in the papers we reproduced.

Keywords: reproducibility, Registered Reports, data sharing, open science
Word count: 4926

Researchers are currently exploring ways to make science more open and transparent. Among novel developments such as pre-registration, preprints, and open peer review, 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 others to more easily reproduce, check, and build on existing work. This requires the development of new skills and best practices, since 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 computationally reproduced the main results of Registered Reports published in the psychology literature.

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

Correspondence concerning this article should be addressed to Daniel Lakens, ATLAS 9.402, 5600 MB, Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl

It is desirable that research is reproducible. Data availability has the potential to make science more efficient by facilitating the re-use of data. The availability of analysis code makes it possible for peers to check and correct 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." 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 link 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 scientific manuscript should be reproducible, which means the data and code can be used to compute the results that are reported in the published article.

Recently, scholars have started to empirically examine the extent to which data is shared with published articles, and, if so, whether it was possible to reproduce the data analyses

reported in the articles. Hardwicke et al. (2018) examined the analytic reproducibility of 35 articles published in the journal *Cognition*. Target outcomes that supported the identified substantive finding of eleven articles could be reproduced by independently writing analysis code, without assistance from the original authors, and 13 sets of target outcomes contained at least one outcome that could not be reproduced even with author assistance. The authors estimated that it took between 2-25 hours to reproduce the reported results per article, 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 emailed authors for the code and data, which they received for 71 articles. The results of one article could not be reproduced due to a lack of a software license, and 16 articles' findings could not be reproduced with access to the required software. For the remaining articles, 32 sets of results could be exactly reproduced, 19 could be reproduced with slight differences, and 3 articles yielded significantly different results. Stodden, Seiler, and Ma (2018) analyzed data availability in the journal *Science* in 2011-2012 and found that 26 of 204 (or 13%) of articles provided information to retrieve data and/or code without contacting the authors. For all datasets they acquired after e-mailing authors for data and code, 26% were estimated to be computationally reproducible. These studies reveal that there is clear room for improvement in how reproducible published articles are.

We set out to examine the data availability and reproducibility in 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 to 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 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 (91 Registered Reports had been published as of February 22, 2018), which makes it possible to examine it 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 and/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 manuscript.

The Current Research. Our main objective was to examine how reusable data and code underlying Registered Reports are, based solely on the information provided in those articles. 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 reproducing 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.

Data. All data and the code used to create this manuscript is provided in an OSF repository at https://osf.io/suqz3/. We report all measured variables and conducted analyses.

Method. To find Registered Reports published in psychology, we drew from a database of registered reports maintained by the Center for Open Science.[1] At the start of this project (06-05-2018) this database consisted of 118 published Registered Reports. Seventy-nine articles in this database were published in the psychology literature. We limited our analysis to studies performed by single groups, since such articles are most representative of the researchers currently work, and excluded 7 large scale collaborations where dedicated team members were responsible for making the analysis reproducible, such as Registered Replication Reports (Hagger et al., 2016) and Many Labs projects (Klein et al., 2014; R. A. Klein et al., 2018). Upon further inspection of the 72 papers left in the dataset, it turned out that 10 were not formally Registered Reports. This left 62 studies 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, SPSS, Python, MATLAB, and JASP) and excluded studies that required expertise in software packages we were not trained in (e.g., dedicated EEG software), which led to the exclusion of one additional study.

Results. We set out to reproduce the findings of 62 papers that met our inclusion criteria. For each article, we coded whether the data and code were (a) linked, (b) available, (c) not software-specific, (d) understandable, and (e) reproducible. These five categories were inspired by FAIR data

[1] https://www.zotero.org/groups/479248/osf/items/collectionKey/KEJP68G9

principles that address the findability, accessibility, interoperability, and reusability of data and code, but we did not explicitly code whether papers adhered to the exact definitions of the four FAIR principles (Wilkinson et al., 2016). Our main aim was to examine the reproducibility of results, while adhering to FAIR criteria requires meeting more stringent requirements, for example concerning the presence of metadata (which was missing for all datasets). We considered data and code to be linked when the manuscript included a unique link to the data and 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 papers, 45 linked to data and/or code that were part of the article. Linking to data or code does not mean the data is actually available, and not linking to data in an article does not mean data is unavailable. For one article the link no longer worked (highlighting the benefit of using a stable DOI for linking to data). For three articles, the link still worked, but there was no data at the linked destination. For three articles where there was no link to the data in the article, but we were nevertheless able to find data on the Open Science Framework when searching for the title of the paper. For one paper the data was linked but not available, as the data was embargoed until a future date.

In line with our predictions that authors of Registered Reports would be relatively likely to also adopt other open science practices, such as data and code sharing, 43 out of 62 articles in our final dataset shared at least some of the 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 of 174 articles (49.0%) had reusable data, and only 18 out of 174 articles (10.3%) provided the analysis code (Hardwicke et al., 2018). For 37 articles, the available code contained the statistical analyses, and for 41, all the required data files were available to reproduce the reported results (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 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 the code is annotated well enough to be recoded in R. Note that we had access to SPSS and MATLAB and therefore reproduced the analyses using the original .sps scripts where available. When examining whether analyses were reproducible, we only used the same software packages as had been used by the original authors. The data files of one article where an EEG study was reported 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 to share data 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 the dataset variables in a codebook. If the variables are not clearly described (e.g., the dataset consists of variables identified by abbreviations that only make sense 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 papers with available data in a format that was not software-specific, only 24 datasets were described in enough detail to be understandable. This highlights the importance of adding a codebook with a datafile.[2]

Finally, we examined how many of the 36 articles with data and code 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 coded SPSS, R, MATLAB, Python, and JASP analyses regarding (a) the executability of the script, and (b) the reproducibility of the results. After the initial coding, inter-rater 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 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 coders sometimes took different approaches to how much they would adjust the code to make it run on the data (for detailed comments, see our data file). 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 when attempting to run all the code,

[2]For an explanation of how to create machine-readable codebooks (which was also used to create the codebook that is part of this manuscript), see Arslan (2019).

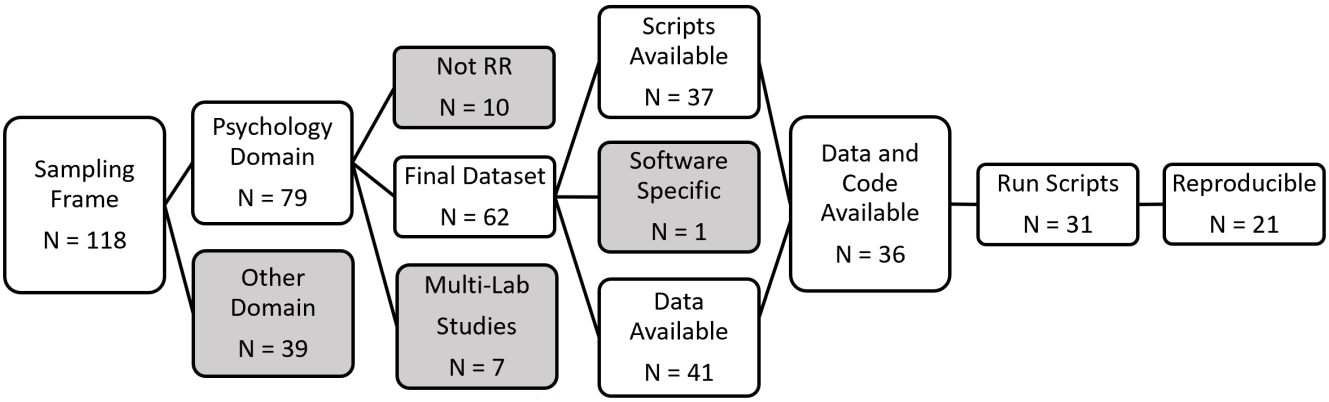

Figure 1. Of all articles in the Registered Report database 79 were psychology related, and 62 were registered reports not part of a multi-lab Registered Replication Report. Of these 62 papers, 45 linked to the data or code. After searching the OSF, data or code was available for 43 studies, of which 39 articles used SPSS, R, MATLAB, Python, or JASP. Of the 62 papers 37 shared the analysis script and 41 shared all data. The analysis script could be run for 31 articles. For the 36 articles for which data and code were available, the main results could be reproduced for 21.

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 coders initially used different thresholds of reproducibility, based on whether every single result reported in a paper could be reproduced by the code and data, or only the main results reported in the article. After evaluating our initial coding round, we considered an article reproducible when we could get the same *main results* as represented in the paper with at best *minor changes* to the analysis scripts. This means we considered the analysis reproducible, even if in the absence of a codebook coders had to search through the analysis code and dataset to identify how variable and label names related to the results reported in the paper. Furthermore, we changed folder locations when needed, and installed and/or 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 the article we identified main results (i.e., any reported descriptives and statistical tests) based on the research question highlighted in the title and abstract. Approximately half of the Registered Reports consisted of replication studies, where the main analysis was explicitly stated based on a previous study. In the remaining Registered Reports the main research question was often stated clearly in a "confirmatory analyses" section. Nevertheless, some arbitrary judgments were required when deciding which analyses were part of the main results.

There were differences between coders in how much expertise they had with R, SPSS, and JASP (PO had less experience, while JG, NC and DL had more experience; SG ran all MATLAB and Python code). The more expertise coders had, the easier it was to reproduce findings (which lead to lower inter-rater reliability). This raises the question regarding which level of skill is expected to be able to reproduce results reported in a scientific article. The required experience might be difficult to quantify. Our results concerning the reproducibility of results is based on which results a PhD student with experience in the statistical software and educated in the same scientific discipline could reproduce, which we feel is a reasonable standard. For the current analysis, we considered a study reproducible as long as the more experienced coder could reproduce the main results. All authors collectively discussed disagreements, which on several occasions led to clarifying ambiguities in rating strategies and correcting mistakes in the process of paper analysis (e.g. overlooked script files). The final ratings presented here have been discussed and approved by all authors, but will not be completely flawless, and we expect different teams of coders to 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 is to evaluate where there is room for improvement, and identify practices that researchers can use to make their work more reproducible.

For each article, two coders examined whether the code could be run on the data, and for five articles where some uncertainty remained, a third coder attempted to reproduce the results. Of the original 62 articles, the analysis code and data were available for 36 articles. We were able to run the code for 31 out of the 37 articles. R was used as a coding language in 13, of which 10 scripts could be run on the data, 17 papers used SPSS, for 15 of which the script could be run, and 3 papers used both SPSS and R, for which the scripts could be run for all 3 papers. 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, the 2 articles that relied on JASP for the analyses always allowed us to reproduce the analyses. Being able to run the code on the data does not imply that all the main analyses reported in the manuscript are correctly reproduced. Some analyses in a paper might not be part of the code or the output. We found that 21 out of 36 papers (58.30%) that included their data and code could be used to reproduce the main results in the article.

For the 15 articles that could not be reproduced, the main reasons were that 1) code to reproduce some values was missing (e.g., dedicated code to run a macro in SPSS), which occurred 8 times, 2) code gave errors (e.g., because variables in the dataset were missing, or because functions did not run as expected), which occurred 6 times, and 3) in one case the code might have been reproducible, but the code was so complex that after 40 minutes only a small part of the results could be reproduced, and it was judged that the time needed to reproduce the results fell outside of a reasonable time span. In the articles that were coded as reproducible, two small errors were observed (one value of 0.89 was rounded to 0.88, and probable type where a value of 0.06 should have been 0.03) but since dozens of other values in these articles 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 were identified by looking through the article, the average time to reproduce analyses reported in R was 27.08 minutes (SD = 28.55) for the first coder and 32.50 minutes (SD = 20.95) for the second coder. Reproducing the SPSS analysis took on average 17.35 minutes (SD = 9.54) for the first coder and 25.50 minutes (SD = 9.72) for the second coder. Most of the time was spent on matching output from the statistical analyses to the analyses reported in the manuscript. This suggests that even if results are reproducible, the organization of the output, and the relation of the output to the published manuscript, can often be improved.

Discussion. We analyzed data and scripts of 62 Registered Reports to examine how many authors shared their data and code, and how often main results reported in the manuscript could be reproduced. In total, 36 out of 62 (or 58.10%) of the articles 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 to authors of articles in political science (Stockemer et al., 2018). Compared to authors of non-registered reports in psychology, data and code sharing is also relatively high, as is the reproducibility rate (taking into account that we reproduced articles without contacting the original authors). The reproducibility rate was higher than the 31% rate observed by Hardwicke et al. (2018), and both data sharing as reproducibility were higher than in the sample of papers from the journal *Science* in 2011-2012 analyzed in Stodden et al. (2018).

Nevertheless, our results indicate that 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). This leads to 4 points researchers in psychology should focus on to improve reproducibility, namely (a) add a codebook to data files, (b) annotate code so it is clear what the code does, and clearly structure code (e.g., using a README) so others know which output analysis code creates, (c) check whether the code you shared still reproduces all analyses after revisions during the peer review process, and (d) list packages that are required in R and the versions used at the top of your R file. We will discuss each of these points below, and link to examples of good practices that we encountered.

First, data is easier to understand and more reusable if variables and their values are clearly described, for example in a codebook. Researchers should ensure that the codebook and variable names are in the same language as the article. Furthermore, when there are multiple datafiles, researchers should provide a clear description of what each datafile contains, for example in a README file in the root directory of the data folder. Le (2018) provides 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 when sharing data.

Second, code should be well-annotated, so that it is understandable for researchers who did not write the code. Well-annotated code makes clear what the analysis code does, in which order scripts should be run if there are multiple scripts (e.g., to pre-process the raw data, compute sum scores, analyze the results, and generate graphs), and which output each section of analysis code generates. A good example of well-annotated code can be found in the materials of Weston and Jackson (2018). It helps to make clear how the analysis code relates to the analyses reported in the paper, to make it easier for others to identify which code generates which results in the paper. For one manuscript which was coded as not reproducible, there was too much unstructured code, and each analyses took too long to run, so that it was decided that the manuscript was not reproducible with a reasonable amount of effort mainly due to the lack of a clear indication which code needed to be run to reproduce specific results. Explicitly linking code in the analysis script to the final manuscript 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 articles

can be found in the materials of Voorspoels, Bartlema, and Vanpaemel (2014). If analyses are performed that are not included in the manuscript, this should be stated explicitly (e.g., assumption checks, exploratory analyses, etc.). 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 scripts). Third, we recommend that researchers perform a final check after peer review has been completed to make sure any changes in the code introduced during the peer-review process are reflected in the shared data and code.

Based on our experiences, we have several specific recommendations for data analyzed in R. First, most code in R relies on specific libraries (also called packages). List all the packages that the code needs to run at the top of the script. Because packages update, it is necessary to report the version numbers of packages that were used (for example using packrat, or copying the output of the sessionInfo() function as a comment in the script). Remember that folder names and folder structures differ between computers, and therefore you should use relative locations (and not "c:/user/myfolder/code"). RStudio projects and the "here" package provide an easy way to use relative paths. When multiple scripts are used in the analysis, include the order in which scripts should be performed on the data in a README file. RMarkdown files provide a useful way to share clearly annotated code and structure the difference steps in the data analysis, for example as done by Campbell et al. (2018).

While trying to reproduce the results of SPSS scripts, the biggest issue was the often confusing and unclear structure of the scripts. Large portion 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 SPSS scripts hindered our attempts to reproduce these results. Often the only time-efficient way to check if an article was 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 script to the analyses reported in the manuscript. Another frequent problem was missing or incorrectly labeled variables in the dataset, so the scripts could not run properly. We expect this is the result of authors updating or modifying either their datasets or their scripts during the publication process. This issue could be easily detected if a second author of the manuscript attempted to reproduce the analyses reported in the final manuscript before data and scripts are shared publicly.

Limitations and Future Research. We limited our analysis to Registered Reports based on the idea that these article formats might be used by people who are early adopters of innovations in science, and would therefore be more likely to also share data and code. We found that the rate at which data and code was shared in our sample was high, relative to studies anlayzed in other reproducibility-focused reviews (e.g, Hardwicke et al., 2018, Stockemer et al., 2018), but we do not have data that gives 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 researchers who published Registered Reports for sharing or not sharing data and code. There are several good reasons why some data should not be shared, and when applicable, researchers should be encouraged to explain their reasons (Morey et al., 2016).

The main aim of this article was not to precisely estimate reproducibility rates, but to see what current standards are, and how the reproducibility of research articles using the Registered Report format could be improved. The sample size is small, and it is doubtful whether a precise estimate of the reproducibility of Registered Reports is of much value, beyond examining where most room for improvement is. 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 issues that can be improved. We 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 above, we believe novel technological solutions might improve the reproducibility of 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 (since the container uses the version of packages 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 (Ragan-Kelley & Willing, 2018), an open-source, browser-based tool for creating and sharing reproducible environments. Another useful technology is

[3]For a Code Ocean capsule reproducing this manuscript, see https://doi.org/10.24433/CO.4275368.v1

RMarkdown, which enable researchers to write fully executable manuscripts. RMarkdown files load the raw data and allow researchers to compute each number reported in the manuscript from the data, instead of copy-pasting values. This means that, as long as the data and require packages can be loaded, all reported numbers can be reproduced. This saves time when matching the analysis code's output to reported results, and thus speeds up the process of checking whether all results reported in the manuscripts are reproduced. The current manuscript is an example of a reproducible RMarkdown 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 results reported in the paper was 24 minutes. This is slightly shorter than the time it took Hardwicke and colleagues (2018), who estimated (without keeping track of the time explicitly) that reproducibility reports took between 2 and 25 person-hours, depending on whether the paper eventually fell in the reproducible or not-reproducible category, and whether author assistance was needed. One major difference between our approaches is that we did not write our own code to analyze the data, as Hardwicke and colleagues did, but simply ran the code written by the original authors on the shared data. We also did not create "reproducibility reports" for each article. Documenting the process of reproducing a paper adds transparency and allows others to check the the decisions about every value in an article. Whether such a level of detail is worth the additional time investment of documenting each reported value is a cost-benefit analysis that journals should make 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 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 reproducibility of manuscripts (i.e., where one checks whether the *main* results in the paper are reproduced by the analysis scripts, but without documenting this step at the level of each individual number) is a surmountable hurdle for journals, and would substantially improve the computational reproducibility of the published literature.

In addition to novel technologies, most progress can probably be made by developing standards within research communities, and educating 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 crystalize, standards should emerge that improve reproducibility, and that 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.

Author Contributions

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

Conflict of Interest Statement

One author (SG) worked at Code Ocean during the writing of this manuscript. SG joined the project after its accompanying Code Ocean component was already submitted and published, and did not write any of the text concerning Code Ocean.

References

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*, 2515245919838783. doi:10.1177/2515245919838783

Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M. B., Kowalik, K., ... others. (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. (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/Neuroscience2014.1.4

Clyburne-Sherin, A., Fei, X., & Green, S. A. (2018). Computational Reproducibility via Containers in Social Psychology. 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*(1), 166–171. doi:10.1017/S1049096516002353

[4] See https://github.com/Lakens/reproducing_registered_reports/blob/master/reproducing_registered_reports.Rmd

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., ... Zwienenberg, M. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science*, *11*(4), 546–573. doi:10.1177/1745691616652873

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, 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), 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*. Univ 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*(1), 20. doi:10.1525/collabra.158

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., ... Nosek, B. A. (2014). Investigating Variation in Replicability. *Social Psychology*, *45*(3), 142–152. doi:10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., 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*(4), 443–490. doi:10.1177/2515245918810225

Le, B. (2018, May). Open Science Manual. https://bit.ly/2w2F6Xu.

Liu, D., & Salganik, M. (2019). Successes and struggles with computational reproducibility: Lessons from the Fragile Families Challenge. doi:10.31235/osf.io/g3pdb

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., ... others. (2016). The Peer Reviewers' Openness Initiative: Incentivizing open research practices through peer review. *Royal Society Open Science*, *3*(1), 150547.

Nosek, B. A., & Lakens, D. (2014). Registered Reports. *Social Psychology*, *45*(3), 137–141. doi:10.1027/1864-9335/a000192

Ragan-Kelley, B., & Willing, C. (2018). Binder 2.0-reproducible, interactive, sharable environments for science at scale.

Stockemer, D., Koehler, S., & Lentz, T. (2018). Data Access, Transparency, and Replication: New Insights from the Political Behavior Literature. *PS: Political Science & Politics*, *51*(4), 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*, *115*(11), 2584–2589. doi:10.1073/pnas.1708290115

Voorspoels, W., Bartlema, A., & Vanpaemel, W. (2014). Can race really be erased? A pre-registered replication study. *Frontiers in Psychology*, *5*. 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*(3), 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*, 160018. doi:10.1038/sdata.2016.18