Analytic reproducibility in articles receiving open data badges at Psychological Science: An observational assessment

Tom E. Hardwicke[1,2], Manuel Bohn[3], Kyle MacDonald[4], Emily Hembacher[3], Michèle B. Nuijten[5], Benjamin N. Peloquin[3], Benny deMayo[3], Bria Long[3], Erica J. Yoon[3], & Michael C. Frank[3]

[1] University of Amsterdam

[2] Charité – Universitätsmedizin Berlin

[3] Stanford University

[4] University of California, Los Angeles

[5] Tilburg University

Author Note

Tom E. Hardwicke, Department of Psychology, University of Amsterdam & Meta-Research Innovation Center Berlin (METRIC-B), QUEST Center for Transforming Biomedical Research, Charité – Universitätsmedizin Berlin. Manuel Bohn, Emily Hembacher, Benjamin N. Peloquin, Benny deMayo, Bria Long, Erica J. Yoon, Michael C. Frank, Department of Psychology, Stanford University. Kyle MacDonald, Department of Communication, University of California, Los Angeles. Michèle B. Nuijten, Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University.

Correspondence concerning this article should be addressed to Tom E. Hardwicke, Nieuwe Achtergracht 129B, Department of Psychology, University of Amsterdam, 1018 WT Amsterdam, The Netherlands. E-mail: tom.hardwicke@uva.nl

Abstract

For any scientific report, repeating the original analyses upon the original data should yield the original outcomes. We evaluated analytic reproducibility in 25 Psychological Science articles awarded open data badges. Initially, 16 (64%, CI [43,81]) articles contained at least one "major numerical discrepancy" (>10% difference) prompting us to request input from original authors. Ultimately, target values were reproducible without author involvement for 9 (36%, CI [20, 59]) articles; reproducible with author involvement for 6 (24%, CI [8, 47]) articles; not fully reproducible with no substantive author response for 3 (12%, CI [0, 35]) articles; and not fully reproducible despite author involvement for 7 (28%, CI [12, 51]) articles. Overall, 37 major numerical discrepancies remained out of 789 checked values 5% CI [3,6], but original conclusions did not appear affected. Non-reproducibility was primarily caused by unclear analysis reporting. Open data alone is not sufficient to ensure analytic reproducibility.

Keywords: open data, badges, reproducibility, open science, meta-research, journal policy

Analytic reproducibility in articles receiving open data badges at Psychological Science: An observational assessment

```
## Warning: Duplicated column names deduplicated: 'Quote of statement, if provided'
## => 'Quote of statement, if provided_1' [27], 'Page number of statement, if
## provided' => 'Page number of statement, if provided_1' [28]
```

A minimum quality standard expected of all scientific manuscripts is that any numerical values can be reproduced if the original analyses are repeated upon the original data (Bollen et al., 2015). This concept is known as *analytic reproducibility* (LeBel et al., 2018; or relatedly, *computational reproducibility*[1], Stodden et al., 2018). When a number cannot be reproduced, this minimally indicates that the process by which it was calculated has not been sufficiently documented. Non-reproducibility may also indicate that an error has occurred, either during the original calculation or subsequent reporting. Either way, there is a breakdown in the integrity of the analysis pipeline that transforms raw data into reported results. As a result, non-reproducibility can create uncertainty about the provenance and veracity of scientific evidence, potentially undermining the credibility of any associated inferences. As a result, non-reproducibility can undermine data re-use (Hardwicke, Mathur, et al., 2018), complicate replication attempts (Nuijten et al., 2018), and create uncertainty about the provenance and veracity of scientific evidence, potentially undermining the credibility of any associated inferences (LeBel et al., 2018).

Difficulty establishing the analytic reproducibility of research reports has been encountered in several scientific domains, including economics, political science, and

[1] Assessments of computational reproducibility typically attempt to reproduce values by re-running original computational code with the original data and can therefore fail if original code is unavailable or non-functioning [e.g., @stodden2018]. By contrast, assessments of analytic reproducibility typically attempt to reproduce values by repeating the original analysis procedures, which can involve implementing those procedures in new code [e.g., @hardwicke2018a].

psychology (Chang & Li, 2015; Eubank, 2016; Hardwicke, Mathur, et al., 2018; Obels et al., 2019; Stodden et al., 2018). A preliminary obstacle for many such studies is that research data are typically unavailable (Hardwicke, Thibault, et al., 2020; Hardwicke & Ioannidis, 2018; Wicherts et al., 2006). Even when data can be accessed, suboptimal management and inadequate documentation can complicate data reuse (Hardwicke, Mathur, et al., 2018; Kidwell et al., 2016). These failures to adequately preserve and share earlier stages of the research pipeline typically preclude downstream assessment of analytic reproducibility.

A previous study in the domain of psychology largely circumvented data availability issues by capitalizing on a mandatory open data policy introduced at the journal Cognition (Hardwicke, Mathur, et al., 2018). After assessing a sample of 35 articles with available data that had already passed initial quality checks, at least one value could not be reproduced within a 10% margin of error in 24 of the 35 (69%) articles. Reproducibility issues were resolved in 11 of the 24 articles after consultation with original authors yielded additional data or clarified analysis procedures. Ultimately, 64 of 1324 (5%) checked values could not be reproduced despite author involvement. Importantly, there were no clear indications that the reproducibility issues seriously undermined the conclusions of the original articles. Nevertheless, this study highlighted a number of quality control and research transparency issues, including suboptimal data management; unclear, incomplete, or incorrect analysis specification; and reporting errors.

In the present study, we intended to investigate whether the findings of Hardwicke, Mathur, et al. (2018) extended to a corpus of Psychological Science articles that received an "Open Data Badge" to signal data availability. In order to focus specifically on the reproducibility of the analysis process rather than upstream data availability or management practices, we selected only articles that had reusable data according to a previous study (Kidwell et al., 2016). The submission guidelines of Psychological Science specifically stated that authors could earn an Open Data badge for "making publicly

available the digitally-shareable data necessary to reproduce the reported result."[2] Additionally, authors were asked to self-certify that they had provided "...sufficient information for an independent researcher to reproduce the reported results."[3] Thus, if this policy was operating as intended, all numbers presented in these articles should be independently reproducible. If not, we hoped to learn about causes of non-reproducibility in order to identify areas for improvement in journal policy and research practice.

Methods

The study protocol (hypotheses, methods, and analysis plan) was pre-registered on October 18th, 2017 (https://osf.io/2cnkq/). All deviations from this protocol or additional exploratory analyses are explicitly acknowledged in Supplementary Information D.

Design

We employed a retrospective non-comparative case-study design based on Hardwicke, Mathur, et al. (2018). The numerical discrepancy between "old" values reported in the target articles and "new" values obtained in our analyses was quantified using percentage error[4] ($PE = \frac{|new-old|}{old} \times 100$). Numerical discrepancies were classified as "minor" ($0\% > \text{PE} < 10\%$) or "major" ($\text{PE} >= 10\%$). If an old p value fell on the opposite side of the $\alpha = .05$ boundary relative to a new p value we additionally recorded a "decision error". We recorded if there was insufficient information to complete aspects of the analyses.

[2] See https://perma.cc/SFV8-DAZ6 (originally retrieved 9th October, 2017).

[3] See https://perma.cc/N8K7-DXP9?type=image (originally retrieved 9th October, 2017).

[4] An important caveat of this measure should be noted: large percentage differences can occur for small absolute differences when the absolute magnitude of values is small (and vice versa). Thus, when considering individual cases, quantitative measures should be evaluated in the full context of our case-by-case qualitative observations (Supplementary Information E).

We classified articles as "not fully reproducible" (any major numerical discrepancies, decision errors, or insufficient information to proceed) or "reproducible" (no numerical discrepancies or minor numerical discrepancies only) and noted whether author involvement was provided or not. We recorded the potential causal loci of non-reproducibility and judged the likelihood that non-reproducibility impacted conclusions drawn in the original articles. Team members provided a subjective estimate of the time they spent on each reproducibility check. Additional procedural details are available in Supplementary Information @ref(sup_design).

Sample

The sample was based on a corpus of psychology articles that had been examined in a previous project investigating the impact of an "open badges" scheme introduced at Psychological Science (Kidwell et al., 2016). This sample was selected because the open badges scheme and assessment of reusability by Kidwell et al. (2016) enabled us to largely circumvent upstream issues related to data availability and data management and instead focus on downstream issues related to analytic reproducibility. A precision analysis indicates that the sample size affords adequate precision for the purposes of gauging policy compliance (Supplementary Information @ref(sup_sample)).

Of 47 articles marked with an open data badge, Kidwell and colleagues had identified 35 with datasets that met four reusability criteria (accessible, correct, complete, and understandable). For each of these articles, one investigator (TEH) attempted to identify a coherent set of descriptive and inferential statistics, roughly 2-3 paragraphs of text, sometimes including a table or figure, related to a "substantive" finding based on "relatively straightforward" analyses. In total, 789 discrete numerical values reported in 25 articles were identified as target values. The articles were published between January 2014 and May 2015. Further information about the sample is available in Supplementary Information @ref(sup_sample).

Procedure

The primary aim of the reproducibility checks was to recover the reported target values by repeating the original analysis (as described in the original articles and any additional documentation such as supplementary materials, codebook, analysis scripts) upon the available data. Attempting alternative analyses was outside the scope of the study. To minimize error and facilitate reproducibility, each reproducibility check was conducted by at least two team members who documented the reanalysis process in an R Markdown report (available at https://osf.io/hf9jy/). If articles were initially classified as not fully reproducible, we emailed the author(s) of the article and used any additional information they provided to try and resolve the issues before a final classification was determined. Additional procedural details are available in Supplementary Information @ref(sup_design).

Data analysis

The results can be considered at several layers of granularity. Detailed qualitative information about each reproducibility check is available in the individual reproducibility reports (https://osf.io/hf9jy/) and summarized in a short "vignette" available in Supplementary Information E. We report descriptive and inferential statistics at the article-level and value-level. We report descriptive and inferential[5] statistics at the article-level and value-level. Ninety-five per cent confidence intervals (CIs) are displayed in square brackets.

[5] Note that although one goal of the study was to characterize analytic reproducibility within a finite sample as a follow-up to Kidwell et al. (2016), we were also interested in generalizing to other psychology articles. To address this second goal, we use standard inferential statistics, recognizing that, because they are generated from a convenience sample, the generality of our findings may be limited.

Results

Prior to author involvement, our reproducibility checks indicated that 16 out of the 25 (64%, CI [43,81]) articles contained at least one major numerical discrepancy. After requesting input from original authors, several issues were resolved through the provision of additional information[6] leaving 10 out of the 25 (40%, CI [22,61]) articles containing at least one major numerical discrepancy. Even after author contact, we could not complete 3 (12%, CI [3,32]) reproducibility checks because insufficient information was available about aspects of the original analyses.

Ultimately, target values were fully reproducible for 9 (36%, CI [20, 59]) articles without requiring author involvement; fully reproducible for 6 (24%, CI [8, 47]) articles only with author involvement; not fully reproducible for 7 (28%, CI [12, 51]) articles despite author involvement; and not fully reproducible for 3 (12%, CI [0, 35]) articles with no substantive response from authors. Availability of original analysis scripts appeared to provide some modest benefits in terms of reproducibility outcomes and time expenditure (see Supplementary Information @ref(sup_results)).

In no cases did the observed numerical discrepancies appear to be consequential for the conclusions stated in the original articles (see Supplementary Information E).

After author involvement, 37 major numerical discrepancies remained amongst the 789 target values examined across all articles (5% CI [3,6]). This included 1 decision error for which we obtained a statically significant p value of 0.02 in contrast to the reported non-significant p-value of 0.09. A scatterplot illustrating the consistency of old and new p-values is displayed in Figure 1. The frequency of numerical discrepancies by value types is displayed in Figure 2.

---

[6] All except one case involved the provision of additional information that was not included in the original article or clarification of original information that was ambiguous or incomplete.In the exception case (4-1-2015_PS), the authors pointed out that we had missed a relevant footnote in the original article.

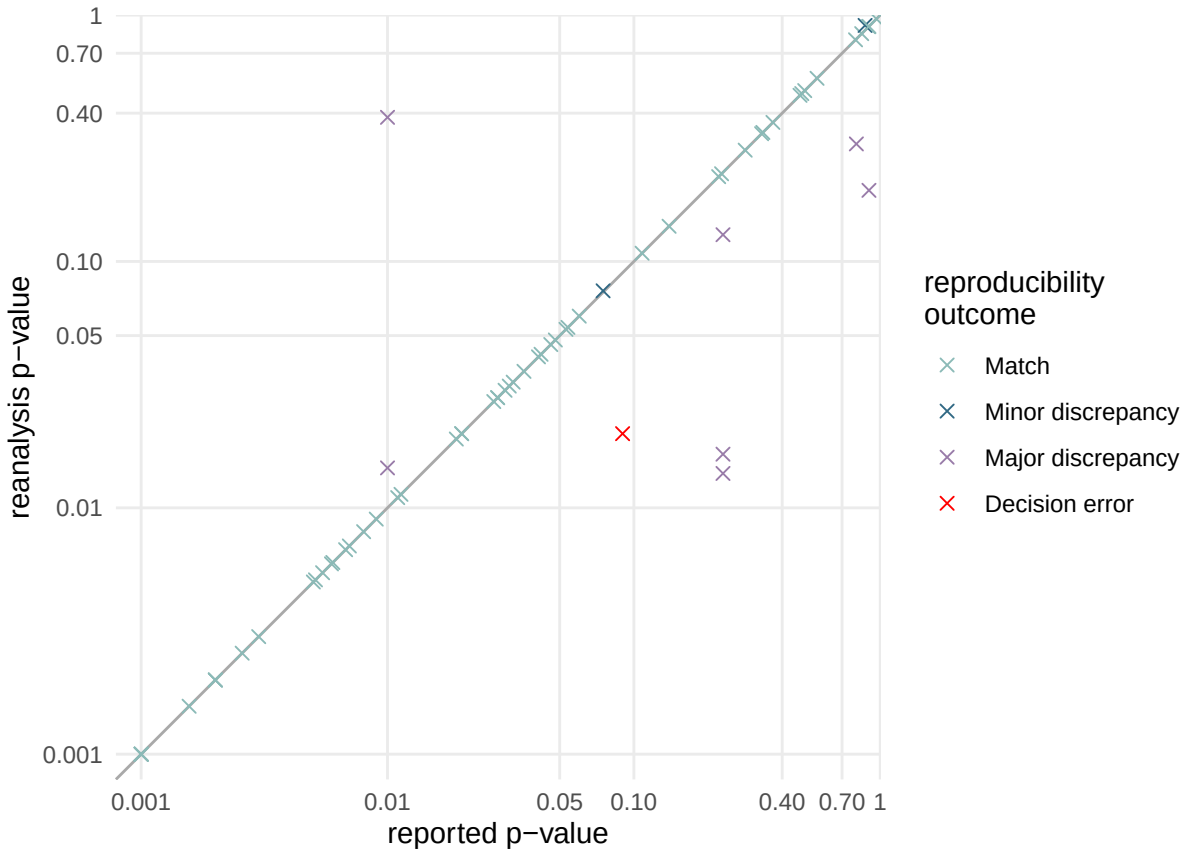

Figure 1. Scatterplot showing reported p-values as a function of reanalysis p-values, classified by reproducibility outcome. Axes are on a log scale. For display purposes, 41 values below 0.001 are not shown.

Where possible we attempted to identify the causal locus of the reproducibility issues we encountered though this was not always possible to confirm definitively. Supplementary Figure C1 shows the frequency of four types of discrete causal loci that we determined were involved in non-reproducibility and how many of these issues were resolved through author involvement. The most common issues we encountered were related to unclear, incomplete, or incorrect reporting of analytic procedures. Examples include unidentified statistical tests, unclear aggregation procedures, non-standard p-value reporting, and unreported data exclusions. Most of these issues could be resolved when original authors provided additional information. Less frequently, we encountered typographical errors and some

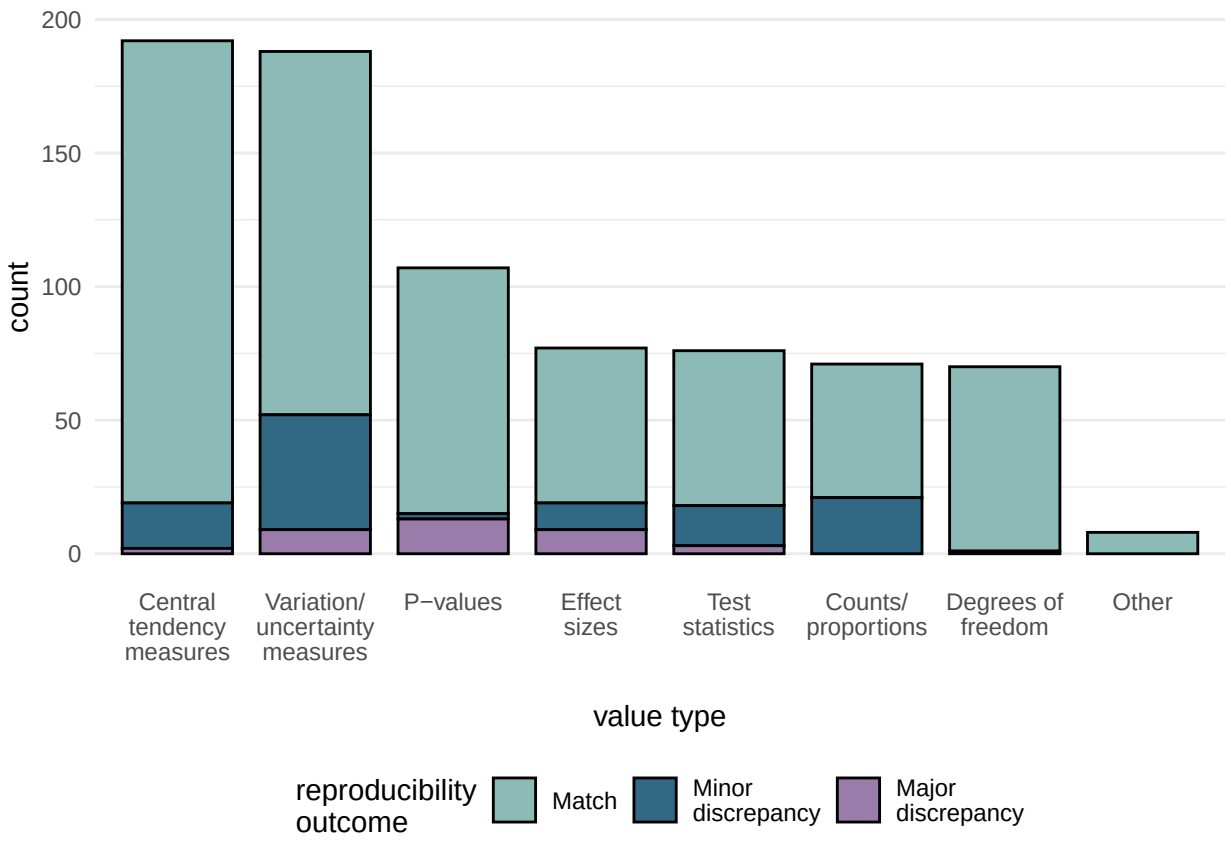

Figure 2. Frequency of reproducibility outcomes by value type. Variation/uncertainty measures include standard deviations, standard errors, and confidence intervals. Effect sizes include Cohen's d, Person's r, partial eta squared, and phi. Test statistics include t, F, and chi squared. Central tendency measures include means and medians.

issues related to data files, including erroneous or missing data. For many instances of non-reproducibility, the causal locus remained unidentified even after contact with original authors.

Team members provided concurrent estimates of their time spent on each stage of the analysis. Altogether, they estimated that they spent between 1 and 30 (median = 7, interquartile range = 5) hours actively working on each reproducibility check (Supplementary Figure C2; total time = 213 hours). This estimate excludes time spent waiting on internal (within our team) and external (with original authors) communications.

Discussion

A reasonable expectation of any scientific manuscript is that repeating the original analyses upon the original data will yield the same quantitative outcomes (Bollen et al., 2015). We have found that this standard of analytic reproducibility was frequently not met in a sample of Psychological Science articles receiving open data badges. Importantly, none of the reproducibility issues we encountered appeared seriously consequential for the conclusions stated in the original articles. Nevertheless, non-reproducibility highlighted fundamental quality control and documentation failures during data management, data analysis, and reporting. Furthermore, the extensive time investment required to check and establish analytic reproducibility would likely be prohibitive for many researchers interested in re-using data. The findings are consistent with a previous study which found a similar rate of non-reproducibility and identified unclear, incomplete, or incorrect analysis reporting as a primary contributing factor (Hardwicke, Mathur, et al., 2018). Although the open badges scheme introduced at Psychological Science has been associated with an increase in data availability (Kidwell et al., 2016), the current findings suggest that additional efforts may be required in order to ensure analytic reproducibility. Although some reproducibility issues were resolved through involvement of original authors, relying on such assistance is not ideal for several reasons: (1) it substantially increases the workload and time investment, both for researchers attempting to re-use data and for original authors; (2) important details about and materials related to the study are more likely to be forgotten or misplaced over time, reducing the ability of authors to assist (Vines et al., 2014); (3) authors may be unwilling or unable to respond, as was the case for three of the articles we examined. Author involvement highlighted that non-reproducibility was often caused by unclear, incomplete, or incorrect reporting of analytic procedures in the original papers, issues which could often be resolved through the provision of additional information or clarifications beyond what was previously reported. In other cases, authors informed us of errors in shared data files or analysis scripts, suggesting that at some stage

the original data, analyses, and research report had become decoupled. For many issues, neither we nor the original authors could reproduce the target values and the causal locus of non-reproducibility remains unidentified; it was no longer possible to reconstruct the original analysis pipeline.

This study has a number of important limitations and caveats. Most pertinently, generalizing the observed reproducibility rates to other articles requires some caution as several characteristics of our sample are likely to have had a positive impact on the reproducibility rate. We specifically selected a sample of articles for which data were already available and screened against several reusability criteria (Kidwell et al., 2016). This is relevant because most psychology articles are not accompanied by raw data or analysis scripts at all (Hardwicke, Thibault, et al., 2020; Hardwicke & Ioannidis, 2018; Wicherts et al., 2006), and even when data is available, suboptimal management and documentation can complicate re-use (Hardwicke, Mathur, et al., 2018). Thus, most psychology articles would likely fail a reproducibility check before reaching the stage of the analysis pipeline at which we began our assessments.

Additionally, all articles in the sample had been submitted to a leading psychology journal that had recently introduced a number of policy changes intended to improve research rigor and transparency (Eich, 2014). These included the open badges scheme, removing word limits for methods and results sections, and requesting explicit disclosure of relevant methodological information like exclusions. It is likely that these new initiatives positively impacted reproducibility, either through directly encouraging adoption of reproducible research practices or attracting researchers to the journal who were already inclined to use such practices. Some evidence, for example, implies that data availability is modestly associated with a lower prevalence of statistical reporting inconsistencies (Wicherts et al., 2011). In sum, several features of the current sample are likely to facilitate reproducibility relative to a more diverse population of psychology articles.

Whilst the current findings are concerning, non-reproducibility is fortunately a solvable problem, at least in principle. As noted above, a major barrier to analytic reproducibility is low availability of research resources like data and code (Hardwicke, Thibault, et al., 2020). Some evidence suggests that journal policies which mandate data sharing can be highly effective at increasing data availability (Hardwicke, Mathur, et al., 2018; Nuijten et al., 2017), apparently more so than voluntary badge schemes (Kidwell et al., 2016). Sharing of analysis scripts may help by providing veridical documentation of the analysis process which is not well captured in verbal prose (Hardwicke, Mathur, et al., 2018). However, as with data sharing, mere availability will not be sufficient to confer the potential benefits of analysis scripts. The utility of scripts depends on a range of factors, including clarity, structure, and documentation, as well the programming language that is used and whether it is in a proprietary format or requires special expertise to understand. In the present study, script availability appeared to offer modest benefits in terms of reproducibility outcomes and time expenditure; however, as only six articles shared scripts[7], it is difficult to draw strong conclusions about their impact from this evidence alone.

Another potential target for policy intervention includes improving quality control at the journal level, potentially through independent assessment of analytic reproducibility; however, this would naturally require additional resources (for discussion see Sakaluk et al., 2014). Some journals, such as the American Journal of Political Science, pay for a third-party institute to verify the reproducibility of all manuscripts prior to publication (Jacoby et al., 2017). Ideally, initiatives intended to improve analytic reproducibility should undergo empirical scrutiny in order to evaluate costs and benefits and identify any policy shortfalls or unintended consequences (Hardwicke, Serghiou, et al., 2020).

Although journal policy is potentially helpful for incentivisation and verification, the

[7] Note that Psychological Science's author guidelines explicitly state that the criteria for an open data badge includes sharing "annotated copies of the code or syntax used for all exploratory and principal analyses." See https://perma.cc/SFV8-DAZ6 (originally retrieved 9th October, 2017).

reproducibility of a scientific manuscript is fundamentally within the control of its authors. Fortunately, a variety of tools are now available that allow for the writing of entirely reproducible research reports in which data, analysis code, and research reports are intrinsically linked (e.g., Vuorre & Crump, 2020). Detailed guidance on data sharing, data management, and analytic reproducibility is available to support psychological scientists seeking to improve these practices in their own research (Klein et al., 2018). The present manuscript is an illustration of these practices in action (https://doi.org/10.24433/CO.6557927.v1). It should be noted that ensuring the reproducibility of a scientific manuscript requires a non-trivial time investment. Continued development of user-friendly tools that facilitate reproducibility and dedicated training may help to reduce this burden. Additionally, the costs of ensuring reproducibility could be offset by the benefits of improved workflow efficiency, error detection and mitigation, and potential for data re-use.

Conclusion

This study has highlighted that the analytic reproducibility of published psychology articles cannot be guaranteed. It is inevitable that some scientific manuscripts contain errors and imperfections, as researchers are only human and people make mistakes (Rouder et al., 2019). However, most of the issues we encountered in this study were entirely avoidable. Data availability alone is insufficient; further action is required to ensure the analytic reproducibility of scientific manuscripts.

Open practices statement

The study protocol (hypotheses, methods, and analysis plan) was pre-registered on October 18, 2017 (https://osf.io/2cnkq/). All deviations from this protocol or additional exploratory analyses are explicitly acknowledged in Supplementary Information D. All data exclusions and measures conducted during this study are reported. All data

(https://osf.io/qzmpj/), materials (https://osf.io/awxt6/), and analysis scripts (https://osf.io/vwzn4/) related to this study are publicly available on the Open Science Framework. To facilitate reproducibility this manuscript was written by interleaving regular prose and analysis code using knitr (Xie, 2015) and papaja (Aust & Barth, 2020), and is available in a Code Ocean container (https://doi.org/10.24433/CO.6557927.v1) which re-creates the software environment in which the original analyses were performed. Analysis code, reports, and Code Ocean containers are also available for each reproducibility check (https://osf.io/hf9jy/).

Conflict of interest statement

All authors declare that there were no conflicts of interest.

Funding statement

This work received no specific funding. TEH's contribution was enabled by a general support grant awarded to the Meta-Research Innovation Center at Stanford (METRICS) from the Laura and John Arnold Foundation and a grant from the Einstein Foundation and Stiftung Charité awarded to the Meta-Research Innovation Center Berlin (METRIC-B).

Author contributions

TEH and MCF designed the study. TEH, MB, KM, EH, MBN, BNP, BdM, BL, EJY, and MCF conducted reproducibility checks. TEH performed the data analysis. TEH and MCF wrote the manuscript. X provided feedback on the manuscript. All authors gave final approval for publication.

Acknowledgements

We are grateful to the authors of the original articles for their assistance with the reproducibility checks. We thank students from Stanford's Psych 251 class, who contributed to the initial reproducibility checks.

References

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. https://github.com/crsh/papaja

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., & Olds, J. L. (2015). *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*. National Science Foundation.

Chang, A. C., & Li, P. (2015). *Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not"* (pp. 1–26). Board of Governors of the Federal Reserve System.

Eich, E. (2014). Business Not as Usual. *Psychological Science*, *25*(1), 3–6. https://doi.org/10.1177/0956797613512465

Eubank, N. (2016). Lessons from a Decade of Replications at the Quarterly Journal of Political Science. *PS: Political Science & Politics*, *49*(2), 273–276. https://doi.org/10.1017/S1049096516000196

Hardwicke, T. E., & Ioannidis, J. P. A. (2018). Populating the Data Ark: An attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. *PLOS ONE*, *13*(8), e0201856. https://doi.org/10.1371/journal.pone.0201856

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(8), 180448. https://doi.org/10.1098/rsos.180448

Hardwicke, T. E., Serghiou, S., Janiaud, P., Danchev, V., Crüwell, S., Goodman, S. N., &

Ioannidis, J. P. (2020). Calibrating the Scientific Ecosystem Through Meta-Research. *Annual Review of Statistics and Its Application*, *7*(1), 11–37. https://doi.org/10.1146/annurev-statistics-031219-041104

Hardwicke, T. E., Thibault, R. T., Kosie, J., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. (2020). *Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014-2017).* https://doi.org/10.31222/osf.io/9sz2y

Jacoby, W. G., Lafferty-Hess, S., & Christian, T. M. (2017). *Should Journals Be Responsible for Reproducibility? | Inside Higher Ed.* https://www.insidehighered.com/blogs/rethinking-research/should-journals-be-responsible-reproducibility.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology*, *14*(5), e1002456. https://doi.org/10.1371/journal.pbio.1002456

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., Ijzerman, H., Nilsonne, G., Vanpaemel, W., & Frank, M. C. (2018). A Practical Guide for Transparency in Psychological Science. *Collabra: Psychology*, *4*(1), 20. https://doi.org/10.1525/collabra.158

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A Unified Framework to Quantify the Credibility of Scientific Findings: *Advances in Methods and Practices in Psychological Science.* https://doi.org/10.1177/2515245918787489

Nuijten, M. B., Bakker, M., Maassen, E., & Wicherts, J. M. (2018). Verify original results through reanalysis before replicating. *Behavioral and Brain Sciences*, *41.* https://doi.org/10.1017/S0140525X18000791

Nuijten, M. B., Borghuis, J., Veldkamp, C. L. S., Dominguez-Alvarez, L., Van Assen, M.

A. L. M., & Wicherts, J. M. (2017). Journal Data Sharing Policies and Statistical Reporting Inconsistencies in Psychology. *Collabra: Psychology*, *3*(1), 31. https://doi.org/10.1525/collabra.102

Obels, P., Lakens, D., Coles, N., Gottfried, J., & Green, S. A. (2019). *Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology.* https://doi.org/10.31234/osf.io/fk8vh

Rouder, J. N., Haaf, J. M., & Snyder, H. K. (2019). Minimizing Mistakes in Psychological Science. *Advances in Methods and Practices in Psychological Science*, *2*(1), 3–11. https://doi.org/10.1177/2515245918801915

Sakaluk, J., Williams, A., & Biernat, M. (2014). Analytic Review as a Solution to the Misreporting of Statistical Results in Psychological Science. *Perspectives on Psychological Science*, *9*(6), 652–660. https://doi.org/10.1177/1745691614549257

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. https://doi.org/10.1073/pnas.1708290115

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert, K. J., Moore, J.-S., Renaut, S., & Rennison, D. J. (2014). The Availability of Research Data Declines Rapidly with Article Age. *Current Biology*, *24*(1), 94–97. https://doi.org/10.1016/j.cub.2013.11.014

Vuorre, M., & Crump, M. (2020). *Sharing and organizing research products as R packages.* https://doi.org/10.31234/osf.io/jks2u

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. *PLoS ONE*, *6*(11), e26828. https://doi.org/10.1371/journal.pone.0026828

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of

psychological research data for reanalysis. *American Psychologist*, *61*(7), 726–728. https://doi.org/10.1037/0003-066X.61.7.726

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman; Hall/CRC. https://yihui.org/knitr/

Appendix A

Supplementary sample information

We downloaded data from the Kidwell et al. (2016) study (available here: https://osf.io/u6g7t/) and selected the 47 articles that had received an 'Open Data Badge'. Kidwell and colleagues had determined that 35 of these articles were accompanied by data sets that were accessible, correct, complete, and understandable. One team member (TEH) examined each of these articles and attempted to identify a 'substantive finding' supported by a 'relatively straightforward analysis' according to the following operational definitions (adopted directly from Hardwicke et al. 2018): (A) A 'substantive finding' is one that supports a central conclusion of the article under scrutiny. This is to some extent subjective, but generally refers to findings supported in the article's abstract and/or displayed in a figure/table; (B) 'Relatively straightforward analyses' involved behavioural data only and employed quantitative techniques that would be commonly found in an introductory level statistics textbook aimed at undergraduate psychology students (e.g., **???**). Examples included means, medians, standard deviations, confidence intervals, standardized effect sizes, correlations, t-tests, and ANOVAs.

The first finding encountered in the article that met these eligibility criteria was selected. The target values for the reproducibility checks were a coherent set of descriptive and inferential statistics related to this finding, typically 2-3 paragraphs of information, occasionally accompanied by a table or figure. During this assessment, 9 articles were excluded because they did not meet the eligibility criteria (ID codes: 12-8-2014 PS, 21-7-2014 PS, 13-6-2014 PS, 18-7-2014 PS, 3-2-2015 PS, 6-3-2015 PS, 16-8-2014 PS, 19-4-2015 PS, 15-2-2015 PS). An additional article was excluded because the data no longer appeared to be available (contra Kidwell et al., 2016; ID code: 8-11-2014 PS). In total, 25 eligible articles

were identified. A bibliography file is available where article ID codes can be matched to their references (https://osf.io/rszey/). Note that there was a reporting error in our pre-registered protocol which said that article 16-2-2015 was excluded when it was not.

A precision analysis (not pre-registered and performed after study completion) shows the range of expected margin of error (confidence interval width) for the range of possible effect sizes given our sample size of 25 articles (Figure A1). The maximum margin of error is 0.20.

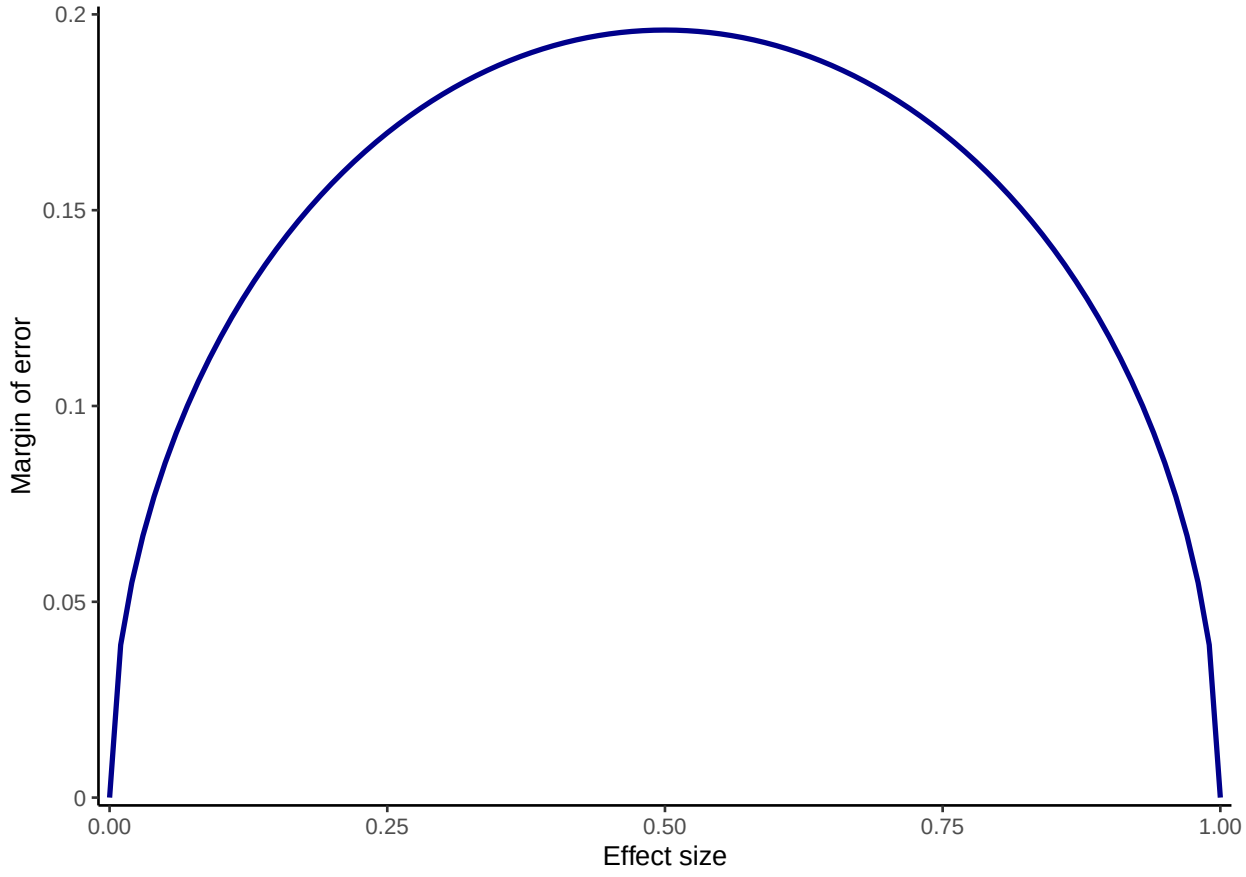

Figure A1. Precision analysis yielding expected margin of error (confidence interval width) relative to effect size (proportion) for a sample size of 25.

Appendix B

Supplementary design and procedures

The most serious consequence of non-reproducibility is that it undermines the credibility of associated scientific claims. We therefore attempted to determine the extent to which any reproducibility issues we encountered had substantial implications for related conclusions drawn in the original papers. As noted in a previous investigation (Hardwicke et al., 2018), this is a complex issue that cannot be straightforwardly measured by a quantitative index. We therefore, considered this question on a case by case basis, using multiple factors to inform our judgement, including the presence/absence of decision errors, the quantity of and type of target values that could not be reproduced, the difference in magnitude of key values like effect sizes, and the specificity of the hypothesis under scrutiny. This was a necessarily subjective exercise; however the judgement for each reproducibility check was agreed upon by the (two or more) team members who worked on it in addition to the first author (TEH) and senior author (MCF). The rationale underlying each judgement is outlined in the reproducibility vignettes in Supplementary Materials E.

Determining the causal locus of non-reproducibility was not straightforward because we often did not have access to comprehensive information about the original analysis process. Additionally, there is not necessarily a direct mapping between discrete causal loci and reproducibility issues; multiple numerical discrepancies could be attributable to a single cause or multiple causes. Sometimes the causal locus of non-reproducibility became apparent after discussions with the original authors, however, all classifications were made independently by our team.

Comparisons between values were partly automated through the use of a custom R function (https://osf.io/teuzb/) that accounted for rounding errors by rounding new values to the same number of decimal places as old values prior to

computing the percentage error and classifying the value as a match, minor discrepancy, or major discrepancy. Note that for 212 values reported relative to a threshold (e.g., $p < 0.05$) or embedded in figures, we could not directly calculate percentage error and it was necessary to 'eyeball' (i.e., visually compare side-by-side) the values to confirm a match.

All team members had started or completed graduate level training in psychology and had experience conducting the types of data analyses typically found in an introductory psychology statistics textbook (e.g., Field et al., 2012).

In the first contact emails we sent to original authors, we explained the project, the issues we had encountered with their article, and provided a copy of an interim reproducibility report. If we were notified that the email was not delivered, we spent up to 5 minutes searching for a more recent contact email address. If the corresponding author(s) did not respond after two weeks, a reminder email was sent.

Appendix C

Supplementary results

Ninety-five per cent confidence intervals (CIs) displayed in square brackets are based on the Wilson method with continuity correction for binomial proportions (**???**) and the Sison–Glaz method for multinomial proportions (**???**).

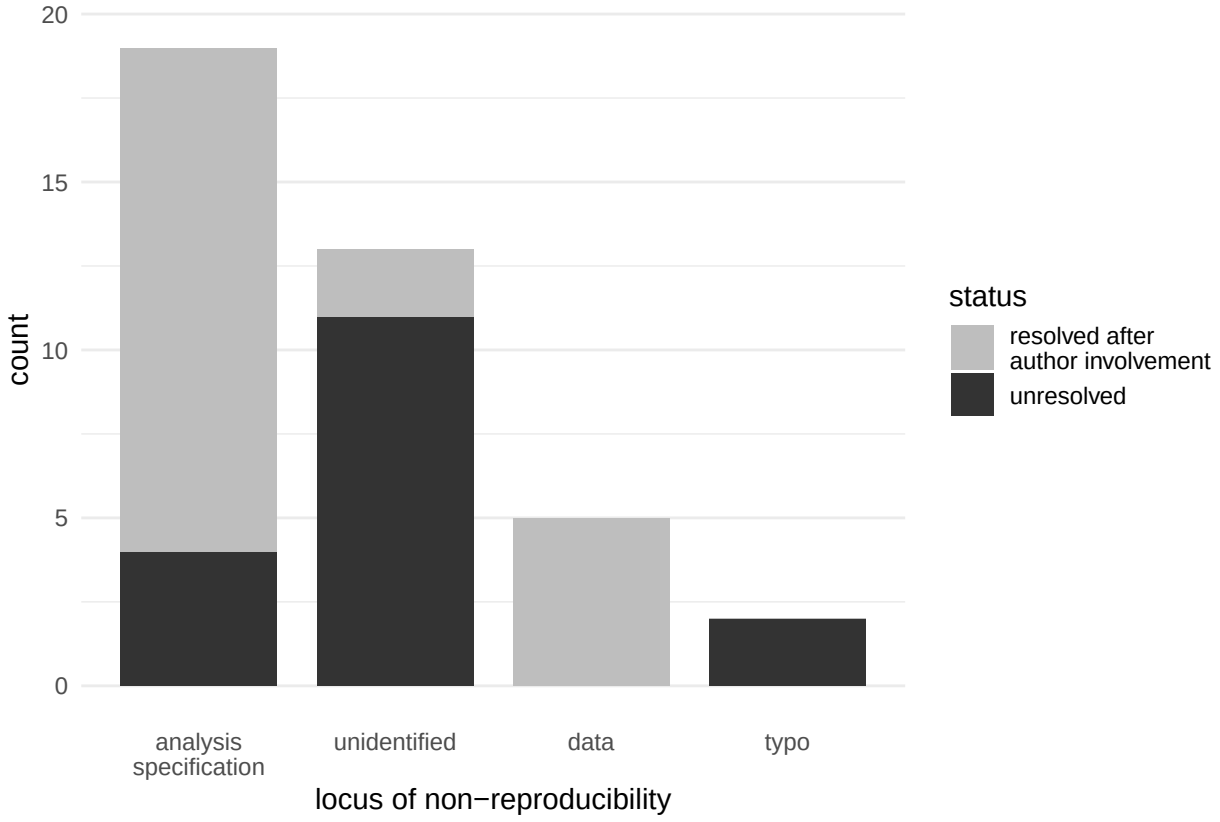

Figure C1. Frequency of discrete causal loci underlying non-reproducibility across all reproducibility checks. Note that some articles contained multiple causal loci.

Six articles were accompanied by analysis code, including one with Matlab scripts and SPSS syntax (8-12-2014_PS), one with Matlab scripts only (2-2-2015_PS), two with SPSS syntax only (2-10-2014_PS, 16-9-2014_PS), one with an R script (8-5-2015_PS), and one with SAS syntax (16-2-2015_PS). Three articles were reproducible without author involvement; one article was reproducible

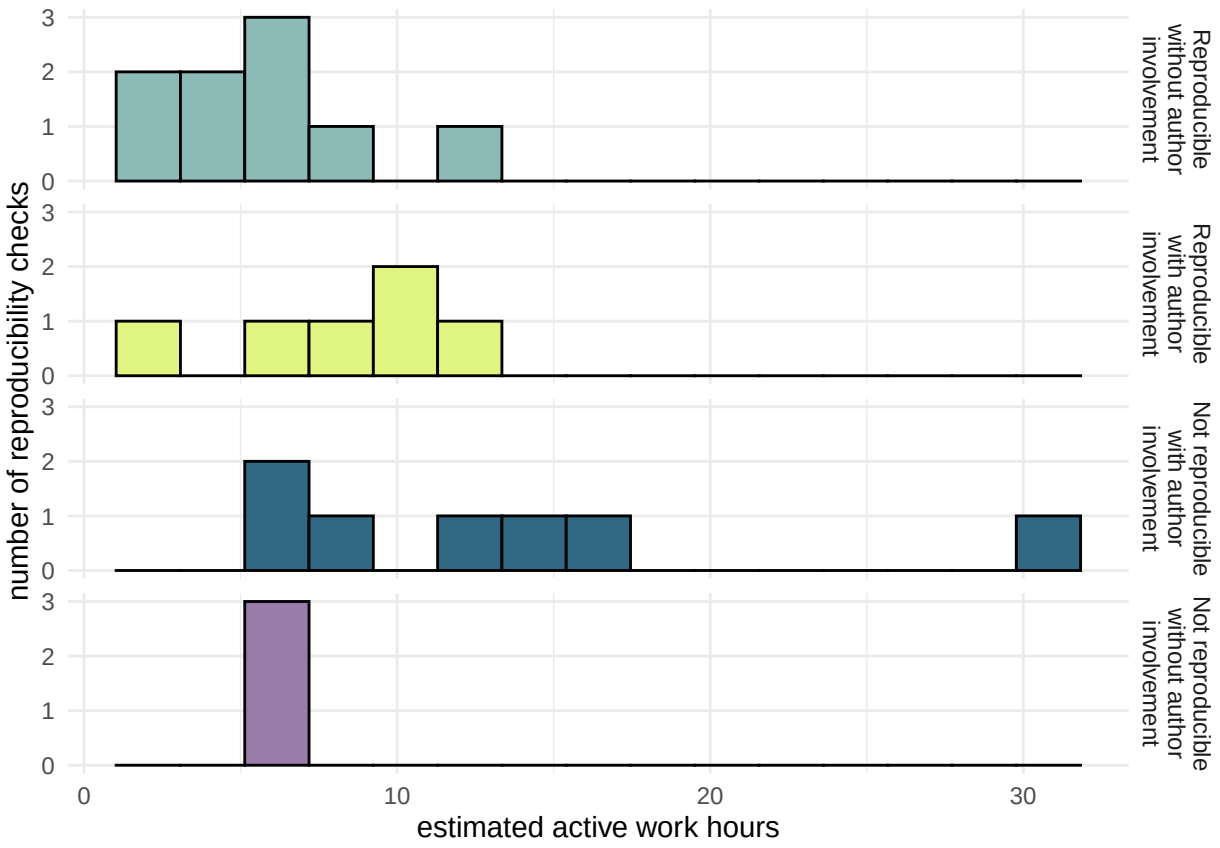

Figure C2. Estimated number of active work hours spent by our team on each reproducibility check, presented separately by reproducibility outcome.

with author involvement; one article was not reproducible with author involvement; and one article was not reproducible with no author response. Team members estimated that they spent between 3 and 10 (median = 6.50, interquartile range = 3) hours actively working on each of the reproducibility checks for articles with scripts.

Appendix D

Deviations from pre-registered protocol

The pre-registered protocol (https://osf.io/2cnkq/) does not state that we will report confidence intervals. However, we decided to report confidence intervals as this would aid inferences beyond the sample.

The pre-registered protocol used the terminology "major/minor errors". To improve clarity, we use "major/minor numerical discrepancies".

The pre-registered protocol used the terminology "reproducibility failure" and "reproducibility success" in reference to article classifications. To improve clarity, we use "not fully reproducible" and "reproducible".

The pre-registered protocol stated that "After contacting authors for assistance, a maximum time-limit of 2 months for resolution of any issues will be observed. Any reproducibility issues that cannot be resolved within this time-period will be considered reproducibility failures for the purposes of the present investigation." We decided not to impose this time limit as it seemed unreasonable that delays on our side could influence reproducibility outcomes. Authors were able to respond until the project was completed on June 2, 2020.

Appendix E

Reproducibility vignettes

Each of the reproducibility checks is summarised below in a short vignette. Each vignette contains a link to a full R Markdown report rendered in HTML that will take the reader through the reanalysis process step-by-step. A link is also provided to an OSF/Github repository that contains all relevant data and analysis code and a link to a Code Ocean container which recreates the software environment in which the original analyses were run to facilitate reproducibility.

Vignette 1 (article 2-2-2015_PS)

Outcome: Reproducible (with author involvement)

Substantial implications for the original conclusions: N/A

R Markdown report: https://perma.cc/L86J-CWUT

OSF/Github repository: https://osf.io/ghst4/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.4497763.v1

Description:

The authors provided extensive Matlab code outlining the original analyses and outputting what appeared to be the primary results sentences and figures in the paper. However, we found that these did not actually match up with the numbers reported in the original article. We noticed that there were some trial exclusions in the main analysis file that were described as "optional". We tried running the code both when excluding these optional trials and when including them, however, neither matched the values reported in the paper.

We asked the authors for their input. It turns out that they had already

realised the original code they uploaded did not completely reproduce the reported findings, and had additionally discovered a mistake in the data they had posted to the Open Science Framework. They had already posted corrected data and code to the OSF (https://osf.io/mx92g/) but our team was not aware of this as it was not mentioned in the paper. The authors say that they had considered posting a corrigendum to the paper but decided not to pursue that because Psychological Science only publishes one if the error "significantly affects an article's findings or conclusions or a reader's understanding." The authors "did not feel this met that threshold".

A readme file accompanying the updated OSF materials contains further details about the original errors:

1) The "Example analysis code" component uploaded to OSF 10/17/2014 (https://osf.io/7gvh9/) was not sufficiently documented and hard to use in replicating the published results. The code presented here is a commented version of the code used to produce the reported analyses, and therefore identically replications the results from the paper.

2) The data file with the mouse traces (sullivanEtAl2014ForPub_mouse.csv) contained several errors, leading to another source of discrepancy from the paper's reported statistics. The mouse tracking data uploaded in this component (newData.csv) is the original raw data.

3) This code reports that the percent of trials excluded from analysis is 8.1%, but the paper reports that it was 13.3%. This updated 8.1% figure is correct, and the reported number in the paper is a mistake."

With the updated data and code, we are now able to reproduce the values and

have attributed the original reproducibility problems to the errors in the data file as specified above by the authors.

Vignette 2 (article 3-10-2014_PS)

Outcome: Not fully reproducible (with author involvement)

Substantial implications for the original conclusions: Unknown (analysis could not be completed)

R Markdown report: https://perma.cc/6FBE-GXE6

OSF/Github repository: https://osf.io/9vqs5/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.7978743.v1

Description:

We were able to reproduce the descriptive results presented in Figure 2 but encountered problems reproducing the inferential statistics. After receiving a more specific analysis specification from the author, we were mostly able to reproduce the key 3-way interaction from the repeated measures ANOVA. Specifically, we were able to reproduce the degrees of freedom and the p-value. However, we still found a minor numerical discrepancy in the F-statistic and a major numerical discrepancy with the η^2 (although the absolute magnitude of the difference was small).

Despite several attempts, we were only able to reproduce a subset of the follow-up Tukey tests and needed further information to proceed. One p-value appeared to be a decision error, because it was reported as $p > 0.09$ but we obtained $p = 0.02$.

We requested the following additional information from the authors: (1) the calculation that they used for the η^2 value in the repeated measures ANOVA; (2)

precisely what subsets of the data were used for each p-value reported in the follow-up Tukey tests. However, we did not receive a response to this request and have concluded that there is insufficient information to proceed with the analysis.

Vignette 3 (article 1-1-2015_PS)

Outcome: Not fully reproducible (without author involvement)

Substantial implications for the original conclusions: Unlikely.

R Markdown report: https://perma.cc/MY9V-97XQ

OSF/Github repository: https://osf.io/6p28a/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.7037066.v1

Description:

We could reproduce all target values aside from one major numerical discrepancy for an F-value. In the results section of the paper it was stated that all F-values of the non-significant tests were <1, but in model 3, the F-value we obtained was 2.21. We contacted the first author on Apr 25 and both the first author and last author on May 22 to ask for their input. We've received no response and have therefore concluded this reproducibility check. Although the cause of non-reproducibility cannot be identified, this single observed discrepancy is unlikely to be consequential for the article's original conclusions.

Vignette 4 (article 11-11-2014_PS)

Outcome: Not fully reproducible (with author involvement)

Substantial implications for the original conclusions: Unlikely

R Markdown report: https://perma.cc/K2Y6-58YK

OSF/Github repository: https://osf.io/huqmg/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.3190579.v1

Description:

This reproducibility check was largely a success. There were some difficulties mapping the reported analyses to the variable names in the data file, however these were eventually resolved with a bit of guesswork.

There was one major numerical discrepancy where a p value was reported as = .001 but we obtained 3.695211e-13. It seems very likely the intention was to report p <.001 as the reported t-value is consistent with this significance level (the authors have confirmed over email that it was probably a typo). This single observed discrepancy is unlikely to be consequential for the article's original conclusions.

Vignette 5 (article 11-12-2014_PS)

Outcome: Reproducible (without author involvement)

Substantial implications for the original conclusions: N/A

R Markdown report: https://perma.cc/29RS-Y55L

OSF/Github repository: https://osf.io/btpnm/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.5508100.v1

Description:

We encountered no major difficulties reproducing all target values for this article.

608 **Vignette 6 (article 16-11-2014_PS)**

609 Outcome: Reproducible (without author involvement)

610 Substantial implications for the original conclusions: N/A

611 R Markdown report: https://perma.cc/USK3-244C

612 OSF/Github repository: https://osf.io/ubwjm/

613 Code Ocean reproducible analysis container:
614 https://doi.org/10.24433/CO.2068771.v1

615 Description:

616 The inferential statistical tests employed were not explicitly named in the
617 original article. However, we made an educated guess that they were chi-squared
618 tests given the context, and this enabled us to reproduce all target values.

619 **Vignette 7 (article 1-6-2014_PS)**

620 Outcome: Not fully reproducible (with author involvement)

621 Substantial implications for the original conclusions: Unlikely.

622 R Markdown report: https://perma.cc/89A9-QLXN

623 OSF/Github repository: https://osf.io/4k6v5/

624 Code Ocean reproducible analysis container:
625 https://doi.org/10.24433/CO.5582627.v1

626 Description:

627 Note that for this reproducibility check, a corrigendum had previously been
628 published for the target article, so we aimed to reproduce the values reported in the
629 corrigendum rather than the original article. We also could not work out how to

implement the inferential analyses in R and used SPSS instead - aided by the SPSS syntax provided by the original authors.

All inferential statistics could be reproduced using SPSS. However, we found one major numerical discrepancy in the descriptive statistics (which were run in R): a reported mean of 0.05, which according to our analysis was 0.005. The relevant value also seemed to be 0.005 according to Figure 1 of the corrigendum. We contacted one of the authors and they confirmed 0.005 is the correct value and the value reported in the corrigendum is a typo. This single observed discrepancy is unlikely to be consequential for the article's original conclusions.

Vignette 8 (article 10-7-2014_PS)

Outcome: Reproducible (with author involvement)

Substantial implications for the original conclusions: N/A

R Markdown report: https://perma.cc/53BJ-BP3P

OSF/Github repository: https://osf.io/5eps7/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.4524649.v1

Description:

We could initially reproduce all values except for one p-value reported for a simple effects analysis. After the original authors shared SPSS syntax with us it became clear that they were implementing the analysis in a different way to what we had understood from reading the paper. When we used the original method, we were able to reproduce the p-value successfully.

Vignette 9 (article 16-2-2015_PS)

Outcome: Not fully reproducible (without author involvement)

Substantial implications for the original conclusions: Unlikely.

R Markdown report: https://perma.cc/BD6J-U6V4

OSF/Github repository: https://osf.io/r9j83/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.2045757.v1

Description:

We could successfully reproduce all reported descriptive statistics for the target values; however, we observed several major numerical discrepancies for correlation coefficients and their confidence intervals. However, although these cases exceeded our 10% percentage error threshold, the absolute magnitude of the difference was small (all $< .03$ correlation units). It is not clear to us what the source of the discrepancy is, although it could be that different software packages (the authors used SAS and we are using R) give different values for these tests.

We attempted to run the SAS program which was provided alongside the article. However, the program is trying to read in the following text files: 'explicit.txt' 'sessions.txt', 'sessionTasks.txt', 'iat.txt', which, as far as we can tell, are not shared along with the article. We contacted the corresponding author on April 23, 2019, to request these files. We initially received a response from the corresponding author on Apr 23 (2019) which offered to help in principle but noted that this could take some time. We sent follow-up emails on April 25, May 22, and July 2, but received no further response. We ultimately decided to close the reproducibility check. Note that although we made contact with an author, no substantive assistance was provided - this is reflected in the categorisation of this

case as “not fully reproducible (no author involvement)”

Although we cannot identify the causal locus of non-reproducibility, the magnitude of the differences suggests that none of the issues we identified are likely to substantially change the authors’ original conclusions.

Vignette 10 (article 16-9-2014_PS)

Outcome: Reproducible (without author involvement)

Substantial implications for the original conclusions: N/A

R Markdown report: https://perma.cc/K5LB-WD45s

OSF/Github repository: https://osf.io/hz8mn/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.1533019.v1

Description:

We encountered some numerical differences for sample sizes reported in the original article compared to our reanalysis, however these differences did not meet the threshold to be classified as major numerical discrepancies. We could successfully reproduce all target values.

Vignette 11 (article 3-4-2015_PS)

Outcome: Not fully reproducible (with author involvement)

Substantial implications for the original conclusions:

R Markdown report: https://perma.cc/PGT4-GN8U

OSF/Github repository: https://osf.io/zkmgw/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.6468572.v1

Description:

In this reproducibility check we were able to reproduce the descriptive statistics (Figure 3) but ran into difficulties reproducing some of the inferential statistics. We contacted the authors and received assistance that resolved some issues. Specifically, information was provided about the identity of statistical tests, units of analysis, and normalization procedures, that were not stated in the paper. Unfortunately, we then lost contact with the authors and some reproducibility issues remained that we could not resolve.

We could not reproduce two effect sizes (d). The authors reported that they tried to reproduce these values themselves and could "closely reproduce the Cohen's d values". However, they did not share the exact values they obtained in their re-analysis. We could also not reproduce one mean and one bound of a confidence interval. The causal locus of these issues is unclear.

The reproducibility issues do not appear to undermine the original conclusions. The obtained effect sizes are smaller than those reported, but not substantially so. The mean and ci discrepancies are of low magnitude.

Vignette 12 (article 2-10-2014_PS)

Outcome: Reproducible (without author involvement)

Substantial implications for the original conclusions: N/A

R Markdown report: https://perma.cc/V4R9-V8JJ

OSF/Github repository: https://osf.io/2mq46/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.7548371.v1

Description:

The manuscript did not specify the way that confidence intervals were computed and it was necessary to make an educated guess. All target values were reproduced successfully.

Vignette 13 (article 3-9-2014_PS)

Outcome: Not fully reproducible (with author involvement)

Substantial implications for the original conclusions: Unlikely

R Markdown report: https://perma.cc/4PFB-LHBP

OSF/Github repository: https://osf.io/b3v7h/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.4160541.v1

Description:

Whilst trying to reproduce an ANOVA we noted that two participants (participant 25 for both the Moroccan and Spaniard samples) had missing data for the PAST condition. However, the degrees of freedom reported in the paper appeared to reflect that the participants *were* included in this analysis. We asked the original authors for clarification and they informed us that there were typos in the data file they had posted on OSF. The wrong participant ID had been added for some lines of data, making it appear that two participants were missing data. Once the typos were corrected, the degrees of freedom in our re-analyses matched those reported in the paper.

Two major numerical discrepancies with p-values remained. In both cases, the authors reported that p = .001, but the values we obtained were lower than this. We discussed this issue with the authors and they said that it is their policy to report p values that are <.001 as p = .001. Unfortunately this approach guarantees inaccurate reporting. According to the APA, researchers should "report p values less than .001 as p < .001" (APA, 2009; p. 114). The reproducibility issues we encountered do not appear to undermine the conclusions drawn in the original article.

Vignette 14 (article 4-1-2015_PS)

Outcome: Reproducible (with author involvement)

Substantial implications for the original conclusions: N/A

R Markdown report: https://perma.cc/R2V2-Y3HS

OSF/Github repository: https://osf.io/c6u95/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.4555168.v1

Description:

We initially could not reproduce some degrees of freedom and F-values despite trying multiple different model specifications. We contacted the authors for assistance and they correctly pointed out that we had missed an important footnote (footnote 2) which read: "The sphericity assumption for this analysis and the corresponding analysis in Experiments 2 and 3 was not met. We report multivariate test results, as recommended by Maxwell and Delaney (2004), because these tests are more optimal than correcting for sphericity. The pattern of results did not differ depending on whether we used either a multivariate test or sphericity correction."

They also sent SPSS syntax and a screenshot of the output - which demonstrated successful reproduction of the target values.

Unfortunately at this time our team does not have access to SPSS nor the expertise to independently implement these multivariate tests in R (the analysis appears to be beyond our operational definition of a 'reasonably straightforward analysis'), however the SPSS syntax and output indicates that these outcomes are reproducible.

Vignette 15 (article 4-11-2014_PS)

Outcome: Reproducible (without author involvement)

Substantial implications for the original conclusions: N/A

R Markdown report: https://perma.cc/VA79-D42D

OSF/Github repository: https://osf.io/2kz9b/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.2298575.v1

Description:

We encountered no major difficulties reproducing all target values for this article.

Vignette 16 (article 5-4-2015_PS)

Outcome: Reproducible (with author involvement)

Substantial implications for the original conclusions: N/A

R Markdown report: https://perma.cc/9DEH-7YPY

OSF/Github repository: https://osf.io/bm4cg/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.1310152.v1

Description:

Although it was specified in the paper that one participant was excluded from the analyses due to a computer malfunction, the data files provided online did not specify which participant this referred to. Thus, prior to communication with the author, we were not able to proceed with the analyses. After the author identified the participant to be excluded, we were able to reproduce all target values.

Vignette 17 (article 6-1-2015_PS)

Outcome: Reproducible (without author involvement)

Substantial implications for the original conclusions: N/A

R Markdown report: https://perma.cc/BTS6-5MH6

OSF/Github repository: https://osf.io/f9q28/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.0405521.v1

Description:

We encountered no major difficulties reproducing all target values for this article.

Vignette 18 (article 6-7-2014_PS)

Outcome: Reproducible (without author involvement)

Substantial implications for the original conclusions: N/A

R Markdown report: https://perma.cc/Z8SP-CEDV

OSF/Github repository: https://osf.io/qezax/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.0223565.v1

Description:

We encountered no major difficulties reproducing all target values for this article.

Vignette 19 (article 7-3-2015_PS)

Outcome: Not fully reproducible (without author involvement)

Substantial implications for the original conclusions: Unknown (analysis could not be completed)

R Markdown report: https://perma.cc/77ND-XHKN

OSF/Github repository: https://osf.io/jgp8e/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.7977267.v1

Description:

In our initial attempts, we found a number of minor numerical discrepancies in the descriptive statistics whether we used medians or means. We then encountered major numerical discrepancies in the inferential statistics. There were a number of aspects of the original analysis and data files we were unclear about. Unfortunately despite emailing the original authors several times we have not received a response to our questions (we have received responses saying they will get back to us, but this has not happened $>$ 20 months after the last message). We have thus classified these issues as 'insufficient information errors'. The issues we are unclear about are outlined in full detail in the reproducibility report. In brief, they refer to unclear

labels and codings for variables in the data file, unclear use of means vs. medians, unclear levels of aggregation. Despite trying out multiple combinations of these factors based on educated guesses, we remain unable to reproduce some values and cannot complete parts of the analysis and it is unclear why.

Vignette 20 (article 8-12-2014_PS)

Outcome: Not fully reproducible (with author involvement)

Substantial implications for the original conclusions: Unlikely

R Markdown report: https://perma.cc/R38G-QQKG

OSF/Github repository: https://osf.io/5hv6f/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.0413564.v1

Description:

We initially ran into a problem reproducing the exclusions reported in the paper. We used both the SPSS and Matlab files provided by the authors and attempted to apply the exclusions in several different ways. As all were unsuccessful, we contacted the original authors to ask for assistance. They created a video of themselves performing the analysis in SPSS and reproducing the exclusion numbers reported in the paper. Watching this video helped us to identify that we had misunderstood the sentence in the paper that reported the exclusions, specifically: “Only participants for whom both d’1 and d’2 could be computed for the analysis subset (i.e., who had non-zero counts in every cell) were included; above chance: n = 165, at chance: n = 33.” We had excluded participants with missing values in any of the cells, whereas the original analysis excluded participants with missing values in cells relevant to the analysis subset. After being sent the video, we

were able to reproduce this part of the analysis successfully and proceeded with the remainder of the analysis.

We were unable to reproduce two values - a t-value (0.17 reported vs. 0.12 in our analysis) and a d-value (0.03 reported vs 0.02 in our analysis). We contacted the authors again and they said they also could not reproduce these values either - in their own re-analyses they obtain the same values as we do. The reason for the differences is unknown. The authors suggest it might be due to a change in version of the SPSS software that was used to run the original analysis. Although they qualify as 'major numerical discrepancies according to our operational definition, the magnitude of these differences indicates that they are unlikely to be consequential to the authors' original conclusions.

Vignette 21 (article 8-5-2015_PS)

Outcome: Reproducible (without author involvement)

Substantial implications for the original conclusions: N/A

R Markdown report: https://perma.cc/WN37-9T2K

OSF/Github repository: https://osf.io/uaeqw/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.2646134.v1

Description:

We encountered no major difficulties reproducing all target values for this article.

Vignette 22 (article 8-7-2014_PS)

Outcome: Reproducible (with author involvement)

Substantial implications for the original conclusions: N/A

R Markdown report: https://perma.cc/4779-EB7S

OSF/Github repository: https://osf.io/mxwq7/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.7903145.v1

Description:

Initially we encountered major numerical discrepancies for the analysis of gain scores. The t-value and the effect size obtained were considerably smaller compared to the ones reported in the paper. We originally calculated the gain scores by averaging scores in each test phase (pre and post) in each condition for each subject and then subtracting the averages for the post test from the ones for the pre test. This way of calculating scores was based on the paper's description: "Gain scores were calculated by subtracting each participant's pretest score from his or her posttest score."

We contacted the original authors and they informed us that the gain scores were calculated in a different way, namely by subtracting pre from post test scores for each specific lesson (e.g., plate tectonics) and then averaging across lesson difference scores to get a mean gain score per subject and condition. This also included excluding some lessons for some of the participants due to absence. This way of calculating the gain scores was not obvious based on the information in the paper. The supplementary material also did not include any additional information about gain scores. After correcting the way the gain scores were calculated we were able to reproduce all target values successfully.

Vignette 23 (article 8-8-2014_PS)

Outcome: Reproducible (with author involvement)

Substantial implications for the original conclusions: N/A

R Markdown report: https://perma.cc/58C5-HWDQ

OSF/Github repository: https://osf.io/buegj/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.2241042.v1

Description:

We initially had trouble reproducing the principle components analysis to generate the wise reasoning composite scores from the raw data due to a missing variable for question 4A from Table 1 and 5 missing values in the raw data file that were not mentioned in the paper. But after author clarifications, these issues were resolved.

Some major numerical discrepancies remained, though their absolute magnitude was low. The authors provided SPSS syntax and we were able to implement this and reproduce the values reported in the paper. Despite extensive efforts we were unable to localize the exact reason why we could not reproduce the values in R. Our suspicion is that these might be caused by opaque SPSS/R computational differences. Nevertheless, we were able to reproduce all target values using a combination of R and SPSS.

Vignette 24 (article 9-2-2015_PS)

Outcome: Reproducible (without author involvement)

Substantial implications for the original conclusions: N/A

R Markdown report: https://perma.cc/9VCW-MXVX

OSF/Github repository: https://osf.io/cneqr/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.7685589.v1

Description:

We encountered no major difficulties reproducing all target values for this article.

Vignette 25 (article 9-5-2014_PS)

Outcome: Not fully reproducible (with author involvement)

Substantial implications for the original conclusions: Unknown (analysis could not be completed)

R Markdown report: https://perma.cc/SVH6-8EN6

OSF/Github repository: https://osf.io/fam6g/

Code Ocean reproducible analysis container: https://doi.org/10.24433/CO.8022934.v1

Description:

We were initially able to reproduce the descriptive statistics and the three key figures. But there were four insufficient information errors pertaining to the original analysis specifications that prevented a full reproducibility check (see reproducibility report for details). We contacted the original authors for assistance and they clarified some aspects of the analyses which resolved some issues.

However, we continued to have difficulties with the section reporting post-hoc comparisons. There were three major numerical discrepancies for p-values. The authors report using 'Sidak corrections' but it is unclear if they were applied to the alpha threshold (correctly) or the p-values themselves. If corrections were applied to the p values, then this could potentially explain the differences, but the article does

not identify the family of hypotheses for which corrections were applied. So we also cannot correct the alpha threshold. We have decided to report the major numerical discrepancies for the p values, but also record an insufficient information error as we cannot determine whether these were decision errors or not. Additionally, for one degree of freedom, we obtained 4 whereas the article reports 3. This may be a typo but the cause is not clear.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology*, *14*(5), e1002456. https://doi.org/10.1371/journal.pbio.1002456