AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

© 2020 American Psychological Association
ISSN: 1082-989X

https://doi.org/10.1037/met0000365

The Reproducibility of Statistical Results in Psychological Research: An Investigation Using Unpublished Raw Data

Richard Artner, Thomas Verliefde, Sara Steegen, Sara Gomes, Frits Traets, Francis Tuerlinckx, and Wolf Vanpaemel
Faculty of Psychology and Educational Sciences, KU Leuven

Abstract

We investigated the reproducibility of the major statistical conclusions drawn in 46 articles published in 2012 in three APA journals. After having identified 232 key statistical claims, we tried to reproduce, for each claim, the test statistic, its degrees of freedom, and the corresponding p value, starting from the raw data that were provided by the authors and closely following the Method section in the article. Out of the 232 claims, we were able to successfully reproduce 163 (70%), 18 of which only by deviating from the article's analytical description. Thirteen (7%) of the 185 claims deemed significant by the authors are no longer so. The reproduction successes were often the result of cumbersome and time-consuming trial-and-error work, suggesting that APA style reporting in conjunction with raw data makes numerical verification at least hard, if not impossible. This article discusses the types of mistakes we could identify and the tediousness of our reproduction efforts in the light of a newly developed taxonomy for reproducibility. We then link our findings with other findings of empirical research on this topic, give practical recommendations on how to achieve reproducibility, and discuss the challenges of large-scale reproducibility checks as well as promising ideas that could considerably increase the reproducibility of psychological research.

Translational Abstract

Reproducible findings, that are findings that can be verified by an independent researcher using the same data and repeating the exact same calculations, are a pillar of empirical scientific research. We investigated the reproducibility of the major statistical conclusions drawn in 46 scientific articles from 2012. After having identified over 200 key statistical conclusions drawn in those articles, we tried to reproduce, for each conclusion, the underlying statistical results starting from the raw data that were provided by the authors and closely following the descriptions of the article. We were unable to successfully reproduce the underlying statistical results for almost one third of the identified conclusions. Moreover, around 5% of these conclusions do no longer hold. Successfully reproduced conclusions were often the result of cumbersome and time-consuming trial-and-error work, suggesting that the prevailing reporting style in psychology makes verification of statistical results through an independent reanalysis at least hard, if not impossible. This work discusses the types of mistakes we could identify and the tediousness of our reproduction efforts in the light of a newly developed taxonomy for reproducibility. We then link our findings with other findings of empirical research on this topic, give practical recommendations on how to achieve reproducibility, and discuss the challenges of large-scale reproducibility checks as well as promising ideas that could considerably increase the reproducibility of psychological research.

Keywords: reanalysis, reproducible research, reporting errors, p values, transparency

Richard Artner https://orcid.org/0000-0002-4515-5650
Thomas Verliefde https://orcid.org/0000-0003-3654-3894
Sara Steegen https://orcid.org/0000-0003-3159-5388
Sara Gomes https://orcid.org/0000-0002-2099-6314
Frits Traets https://orcid.org/0000-0003-1376-5363
Francis Tuerlinckx https://orcid.org/0000-0002-1775-7654
Wolf Vanpaemel https://orcid.org/0000-0002-5855-3885

Thomas Verliefde now works at the Department of Psychology, Eberhard Karls Universität Tübingen.

This research was supported by the KU Leuven Research Grant C14/19/054. We thank all authors who kindly shared the raw data of their studies. We also thank Jelte Wicherts for his valuable contributions which helped improve the clarity and precision of this article. Francis Tuerlinckx and Wolf Vanpaemel share last authorship.

The results of this study were presented at the International Meeting of the Psychometric Society (IMPS) in New York (July 2018) in the form of a talk as well as at the workshop Perspectives on Scientific Error in Groningen (October 2018) and at the IOPS Winter conference in Arnhem (December 2018) in the form of a poster.

Correspondence concerning this article should be addressed to Richard Artner, Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102, 3000 Leuven, Belgium. Email: richard.artner@kuleuven.be

Intending to examine the relation between *p* values and Bayes factors in commonly used statistical procedures such as linear regressions and mixed ANOVAs, Vanpaemel et al. (2015) asked for the raw data of 394 articles published in 2012 in four different APA journals: *Emotion*, *Experimental and Clinical Psychopharmacology*, *Journal of Abnormal Psychology*, and *Psychology and Aging*. Their empirical procedure seemed straightforward and simple at the onset: In a first step, identify the key statistical claims for each article where the underlying raw data were provided. Next, redo the original, frequentist analyses underlying these claims to make sure that the correct data was received and that the meaning of the variables is understood, and to verify the identified claims. In a third step, compute an equivalent Bayesian analysis for each claim, using the *BayesFactor* R package (Morey & Rouder, 2015). Finally, compare the frequentist and Bayesian conclusions.

In this article, we are reporting on the first two steps. The reason we believe such a report is worthwhile is that the reproduction of the identified statistical results was seldom straightforward, often utterly frustrating, and, for many articles, impossible. In total, we tried to reproduce 232 key statistical results reached in a selection of 46 articles, making this work the most comprehensive report on reproducibility in psychology to date.

We define the *reproducibility* of a reported statistical result as the ability to be verified by an independent researcher through the act of identification and execution of the underlying (often deterministic) calculation(s) on the same dataset. Further, we define the *replicability* of a research finding as the ability of an independent researcher, collecting new data under sufficiently similar conditions as in the original study, to reach similar conclusions. Under the wide-held assumption that the laws of nature are the same at all times and all places in the universe, replicability constitutes the foundation of the scientific method. Unfortunately, a significant portion of conducted direct replication studies in psychology fails to support the conclusion drawn in the original study (see, e.g., Open Science Collaboration, 2015; R. A. Klein et al., 2018).

From an epistemological viewpoint, reproducibility should be assessed before replicability (and generalizability). After all, it makes little sense to try to replicate (let alone generalize) a finding if the results supporting the finding are numerically incorrect. What complicates matters is that an irreproducible numerical result is not necessarily incorrect. The inability to reproduce the number might be the result of the scientific report lacking fundamental details, or it might be due to a lack of the necessary skills, rigor, or patience on the part of the person trying to reproduce the result. Relative to replicability, little research on reproducibility exists. In particular, it is unclear what proportion of failed replications can be attributed to irreproducible results in the original or the replication study.

Hampering discussions on reproducibility and replicability is the fact that these terms are not universally agreed upon and are often used interchangeably within the scientific literature, leading to a lot of confusion. For example, Freese (2007) discusses the importance of reproducibility in sociology but refers to it as replicability. Xie (2015) uses the term *reproducible research* as the union of reproducibility and replicability in his book about the *knitr* R package which allows the integration of R code into LATEX and Markdown documents. Chang and Li (2015), reanalyzing 59 economics articles using the underlying data and code of analysis provided by the authors, compare their success rate to the success rate of 100 independent replications of psychological experiments (Open Science Collaboration, 2015; which, ironically, sports reproducibility in its title). In the field of economics, until very recently, assessments of reproducibility were almost universally labeled replications or replications in the narrow sense. In the field of psychology, most recent literature uses the terms reproducibility and replicability similarly to us (see, e.g., Epskamp, 2019). An in-depth discussion of terminology in different scientific fields can be found in Goodman et al. (2016). In this article, reproducibility and replicability are always used as defined by us when discussing other articles, regardless of what term was used by the authors.

In the remainder of this article, we start with an overview of studies investigating reproducibility in psychology and other fields. Then, we detail the relevant parts of our empirical investigation, including the selection of articles and key statistical claims and our workflow. Next, we summarize our results, discuss the types of errors found, and dissect the reasons for the encountered difficulties in the reproductions. Building on our observations, we propose a taxonomy for reproducibility, discuss the state of reproducibility in psychology, and give practical recommendations for producing correct and quickly reproducible results in scientific articles. We conclude by highlighting the need for and the challenges of future research on this subject.

Previous Research Assessing Reproducibility

We conducted a Google Scholar search with the following keywords: *analytical reproducibility, computational reproducibility, methods reproducibility, misreporting of results, repeatability, replicability, reporting errors, reproduction of published results*. For each article we found that reported on an empirical analysis of reproducibility in psychology, we looked at all the articles it references and at all articles that cited it. Table 1 lists all published empirical investigations of reproducibility in psychology as of August 2020. To put these findings into perspective, Table 1 further lists important investigations on reproducibility in related sciences. Below, we first discuss the main findings in psychology. We make a distinction between articles that focus on reproducibility in the strict sense, which is based on the raw data and is the topic of this article, and a more loose interpretation of reproducibility, which does not rely on the raw data and could be termed consistency. We then summarize important empirical findings on reproducibility in related fields. Please note that those empirical studies in Table 1 that are based on raw data differ in several relevant aspects and that it is therefore difficult to compare them. As they differ, for instance, in whether code of analysis was used, whether authors were contacted for assistance, the types of statistical models investigated, the definition of what constitutes a successful reproduction, as well as the size of the reproduction team and the effort put into the reproductions rigorous comparisons ought to be done by a detailed reading of the respective articles and supplementary materials.

Checking Reproducibility via Reanalysis of Raw Data

Hardwicke et al. (2018) is the only published investigation in psychology that systematically reproduced statistical results by reanalysis of the underlying raw data. Hardwicke et al. (2018)

Table 1
Empirical Studies on Reproducibility

First author	Year	Scientific field	# of articles	Study design	Findings
Wolins	1962	Psychology	7	Reproduction of “analyses” based on privately shared data.	Three of seven reanalyzed datasets contained severe errors.
Rossi	1987	Psychology	67	Reproduction of test statistics by using summary statistics reported in the articles themselves.	Five out of 21 (Study 1) and 26 out of 46 (Study 2) recalculated test statistics deviated from the reported ones to an extent that could not be explained by rounding. Approximately 13% of tests published as statistically significant in Study 2 were found not significant in the reproduction.
Garcia-Berthou	2004	Multidisciplinary/ Medicine	32	Consistency checks for *p* values.	38% of 32 articles published in Nature and 25% of twelve articles published in *BMJ* contained at least one consistency error. Out of all 244 recomputations, 28 (11.5%) were inconsistent.
Berle	2007	Psychiatry	96	Consistency checks for *p* values.	Percent of inconsistent *p* values in the three analyzed psychiatric journals were 13.9%, 14.8%, and 14.2%, respectively.
Gotzsche	2007	Biomedicine	27	Reproduction of standardized mean differences (SMDs) and their confidence intervals for two randomly selected trials in 27 meta-analyses; reproduction of the complete meta-analysis for those articles where at least one trial could not be reproduced.	Ten meta-analysis had at least one trial where the SMD and/or the length of its confidence interval could not be reproduced. Seven of these had at least one trial where the SMD differed by more than 0.2. The complete reproduction attempt for these 10 meta-analyses found that the pooled estimates and/or the length of their confidence intervals differed by more than 0.1 for seven of them.
Ioannidis	2009	Genetics	18	Reproduction of a randomly selected table or figure in each article via publically available raw data.	Six of the 18 articles could not be reproduced due to a lack of data. Of the remaining 12 articles, four could not be reproduced at all, six were reproduced with minor deviations and only two were reproduced close enough to warrant the label “reproduced in principle”.
Bakker	2011	Psychology	257	Consistency checks for *p* values.	Three-hundred and 94 (9.7%) of 4,077 recalculated statistical results coming from 194 articles were found to be inconsistent. For 50 (1.2%) results the conclusion changed from significant to nonsignificant or vice versa.
Wicherts	2011	Psychology	49	Consistency checks for *p* values.	Consistency checks for 1,148 NHST results with $p < .05$ found that 49 (4.3%) were inconsistent. Forty-seven of the recalculated *p* values were larger than the reported one.
Petrocelli	2013	Psychology	87	Consistency checks for coefficients in single-mediator models.	Thirty-eight (24%) of 156 single-mediator models coming from 87 articles contained inconsistencies that could not be explained by rounding.
Caperos	2013	Psychology	102	Consistency checks for *p* values.	One-hundred and 48 (12%) of 1,212 exactly reported pairs of test statistic and degrees of freedom were found to be incongruent with the reported *p* value. For 28 (2.3%) results the conclusion changed from significant to nonsignificant or vice versa.
Bakker	2014	Psychology	61	Consistency checks for *p* values via *Statcheck.*	Sixty-seven (9.7%) of 886 recalculated statistical results coming from 61 articles published in six journals were found to be inconsistent. For 11 (1.2%) results the conclusion changed from significant to nonsignificant or vice versa.

(*table continues*)

Table 1 (*continued*)

First author	Year	Scientific field	# of articles	Study design	Findings
Veldkamp	2014	Psychology	697	Consistency checks for *p* values via *Statcheck.*	10.6% of 8,105 re-calculated statistical results coming from 697 articles published in six journals were found to be inconsistent. For 0.8% of the results the conclusion changed from significant to nonsignificant or vice versa.
Chang	2015	Economics	67	Reproduction of the "key empirical results" by using publically available raw data and code of analysis.	Twenty-nine of the 67 papers could not be reanalyzed because data or code was not publically available. Of the 38 papers that could be reanalyzed, 22 (58%) were successfully reproduced without contacting the authors, seven (18%) were successfully reproduced after contacting the authors, and nine (24%) contained incorrect data and/or code.
Nuijten	2016	Psychology	30717	Consistency checks for *p* values via *Statcheck.*	10.6% of 258,105 recalculated statistical results coming from 16,695 articles published in eight journals were found to be inconsistent. For 0.8% of the results, the conclusion changed from significant to nonsignificant or vice versa. More than half of the articles included at least one inconsistent *p* value.
Eubank	2016	Political Science	24	Reproduction of all numerical values in the articles by rerunning provided code of analysis on provided raw data files in light of an in-house reproducibility review.	Twenty of the 24 empirical papers needed at least minor modifications in order to run. Fourteen (58%) articles included results that differed from the output of their own code of analysis. Severe errors sometimes required changes in whole columns or tables of results.
Bergh	2017	Strategic Management	88	Reproduction of inferential statistics by using summary and correlational statistics reported in the articles themselves.	Fifty-eight of the 88 articles did not provide sufficient information to reproduce any result. Sixty-two (8.5%) of the recalculated coefficients had a different sign. Fourteen out of the 144 (9.7%) *p* values associated with linear regression, and 12 out of 55 (22%) *p* values associated with SEM, that were reported as significant in the article, were no longer so.
Brown	2017	Psychology	71	Consistency check of summary statistics (mainly the mean) that are based on ordinal data (i.e. Likert scale) by using only data reported in the articles themselves.	A consistency check of 71 articles with the GRIM (granularity-related inconsistency of means) technique found at least one inconsistent mean in 36 (51%) of them. In 16 of those 36 articles, multiple issues were detected.
Naudet	2018	Medicine	17	Reproduction of effect sizes and *p* values for all primary outcomes of randomized control RCTs by using publically or privately shared raw data and, when available, code of analysis provided by the authors.	For 14 (82%) of the 17 reanalyzed RCTs, all identified primary outcomes could be reproduced. One RCT could not be reproduced for a lack of information about the statistical analysis. Two RCTs contained errors but were similar to the reanalyses in terms of magnitude and statistical significance of the effect.
Hardwicke	2018	Psychology	35	Reproduction of a set of inter-related values related to the first identified substantive finding in each article by using publically available raw data and when available code of analysis provided by the authors.	Eleven (31%) articles were reproduced without author assistance, 11 (31%) were reproduced with author assistance, and 13 (37%) were not reproduced despite author assistance.

(*table continues*)

Table 1 (*continued*)

First author	Year	Scientific field	# of articles	Study design	Findings
Maassen	2020	Psychology	33	Reproduction of 500 primary study effect sizes reported and used in 33 published meta-analyses via the information in the corresponding primary study articles.	Two-hundred and 76 (55%) primary study effect sizes could be reproduced without issues. One-hundred and 14 (23%) differed upon recalculation. The remaining effect sizes could not be reproduced because the primary study did not contain sufficient information (54) or because it was unclear which calculations the respective meta-analysis performed (56).

Note. *BMJ* = *British Medical Journal*; NHST = null hypothesis significance testing; SEM = structural equation model.

reanalyzed 35 articles that were published in the journal *Cognition* between 2015 and 2017, a period in which the journal's open data policy required as a prerequisite for publication that the study's raw data are shared on a suitable third party repository. Hardwicke et al. (2018) identified key statistical results coming from basic statistical analyses (e.g., *t* test, ANOVA) for each of the 35 articles under investigation, and systematically reproduced them via R (R Core Team, 2018) using a pilot/copilot approach. In total, 11 out of 35 articles could be reproduced.[1] After contacting the authors of the 24 problematic articles, they were able to reproduce another 11 because code of analysis, additional data sets, or additional information on the calculations of the reported numbers that could not be reproduced were provided. The remaining 13 articles contained identifiable errors and could not be reproduced, despite author assistance. The number of issues in their sample is remarkable insofar that the authors had to share their raw data publically.

Checking Reproducibility Without Raw Data

Without the raw data, reproducibility can be assessed by checking the consistency of reported numbers that are mathematically connected. The first such investigation was conducted by (Rossi, 1987) who recalculated together with his students the test statistics and degrees of freedom for χ^2-tests, independent group *t* tests, and one-way ANOVAs based on the reported means, standard deviations, and sample sizes. Although the sobering results (see Table 1) cannot be considered to be representative of the psychology literature at that time because articles and results were not selected systematically,[2] it does surprise us that this article has only been cited a few times and that its findings did not result in immediate follow-up studies with similar designs.

More recently, two studies found similar results, indicating many inconsistencies in published test statistics. Petrocelli et al. (2013), who cleverly verified the consistency of reported results in single mediator models, and Brown and Heathers (2017), who made use of the fact that summary statistics of ordinal variables (e.g., Likert rating scale), can only take values from a known set of rational numbers which depends on the sample size (see Table 1).

By far the most common type of consistency check in psychology is based on the fact that for fixed degrees of freedom—the parameters that determine the distribution under H_0—the *p* value is a bijective function of the test statistic for *t*, χ^2, and *F* tests. The same holds for *z* test statistics whose distribution is parameter-free. Given data, the random test statistic and its degrees of freedom take specific values and consequently uniquely determine the *p* value. Hence, a potentially incorrectly reported *p* value can be detected without access to raw data by checking whether it matches the reported test statistic and degrees of freedom, an approach which has been pioneered by García-Berthou and Alcaraz (2004). Since then, inconsistencies in published *p* values have been studied in tens of thousands of published psychology articles. Investigations by Berle and Starcevic (2007); Bakker and Wicherts (2011); Wicherts et al. (2011); and Caperos and Pardo Merino (2013) found that between 4.3% and 14.8% of published *p* values are inconsistent. Further investigations by Bakker and Wicherts (2014); Veldkamp et al. (2014); and Nuijten et al. (2016) by means of a text-mining algorithm named *statcheck*, which was developed by Epskamp and Nuijten (2018) and is available both online and in the form of a package in the statistical software R (R Core Team, 2018), found that between 9.7% and 10.6% of published *p* values are inconsistent. Furthermore, between 0.8% and 1.2% of all *p* values were found to be grossly inconsistent, meaning that the statistical conclusion (significance vs. nonsignificance) would change if one were to calculate and interpret the *p* value based on the reported test statistic and degrees of freedom.[3]

A recent empirical study investigated the reproducibility of 33 meta-analyses published in 2011 or 2012 in a variety of journals (Maassen et al., 2020). In total, Maassen et al. (2020) tried to reproduce 500 of the 1,978 primary study effect sizes that were reported and used in the 33 selected meta-analyses via the information in the corresponding primary study articles. They found that only 276 (55%) primary effect sizes could be reproduced without any issues. For 54 (11%) no effect size was recalculated because essential information was missing in the primary study. Further, 114 (23%) effect sizes differed upon recalculation, 52 (10%) of which moderately or large. The remaining 56 (11%)

[1] They consider an article to be reproduced in the absence of any insufficient information errors, any major numerical errors, as well as any decision errors (i.e., no change in significance).

[2] The study was part of a class assignment for students. Rossi (1987) reports to have also recalculated those test statistics that were not selected by the students (144 in total) and that there were relatively fewer reproducibility issues among them.

[3] For an overview of all previous research on *p* value inconsistencies we refer the reader to Nuijten et al. (2016) and their Table 2, and note that they did not indicate that Berle and Starcevic (2007) only reanalyzed exact *p* values.

could only be numerically reproduced by trying out multiple plausible ways of calculating them. Maassen et al. (2020) further investigated the impact of the 114 differing effect sizes on the respective meta-analytic conclusion in various ways and found small or moderate discrepancies in 13 (39%) of them. In none of the 33 meta-analyses did the statistical significance of the pooled effect size change, and no bias toward over- or underestimation could be detected.

Reproducibility in Related Fields

Our literature search revealed two studies in other social sciences and two in the life sciences that systematically used raw data to check the reproducibility of published results. One extensive study in economics by Chang and Li (2015)[4] found problems in 29 out of 67 articles. Especially in the field of economics, reproducibility seems to be a big problem as the study of Chang and Li (2015) was built on worrisome findings in previous empirical assessments of reproducibility in economics research such as Dewald et al. (1986) and McCullough et al. (2006). Two studies, one in genetics (Ioannidis et al., 2009) and one in political science (Eubank, 2016), discovered even greater problems (10 out of 18 and 20 out of 24). A more recent study in medicine (Naudet et al., 2018), on the other hand, found a lesser rate of issues (14 out of 17 could be reproduced; see Table 1).

One approach to checking reproducibility without the raw data in more complex models, which has not yet been used in psychology, is to recalculate *p* values by using not just reported summary statistics (e.g., mean values, standard errors, sample sizes) but also reported inference statistics (e.g., correlations, covariances). An example of such a design is Bergh et al. (2017) who recalculated *p* values in regressions and structural equation models (SEMs) in 30 articles published between 2000 and 2013 in *Strategic Management Journal*. Worryingly, 14 out of the 144 (9.7%) *p* values associated with linear regression, and 12 out of 55 (22%) *p* values associated with SEM that were reported as significant in the article were no longer so in the reanalysis.

The Current Study

Numerically incorrect results unnoticed by the scientific community considerably slow down the accumulation of knowledge, at the very least, as they lead to aggravated uncertainty about what to conclude on the subject at hand. Furthermore, in the case that numerically incorrect results are nonrandom, they systematically distort the inferences. Reporting errors biased toward significance and inflated effect sizes potentially constitute an important factor in the observed disappearance and decline of effects in replication studies such as Open Science Collaboration (2015); Camerer et al. (2016); R. A. Klein et al. (2018), and Camerer et al. (2018). In this article, we report on our systematic investigation of reproducibility via unpublished raw data in psychology.

Method

Selection of Articles and the Key Statistical Claims

The articles that are the focus of our investigation are a subset of the 394 articles published in 2012 in four different APA journals, namely *Emotion*, *Experimental and Clinical Psychopharmacology*, *Journal of Abnormal Psychology*, and *Psychology and Aging*, for which Vanpaemel et al. (2015) asked the corresponding authors for their raw data (see Figure 1) as a basis for the planned comparison between Bayesian and frequentist inference.

Eventually, Vanpaemel et al. (2015) received raw data from 148 of the 394 contacted authors. Before we decided on our final guidelines, we read in the raw data of a handful of articles published in the journals *Psychology and Aging* and *Journal of Abnormal Psychology*. We then started to identify and reproduce key statistical results reported in those articles. This revealed the time-consuming nature of this study and led us to exclude the 72 papers from the journal *Emotion* for which the raw data was shared from our reanalysis. For two articles in *Experimental and Clinical Psychopharmacology*, we only had temporary permission to use the raw data, so these were excluded from reanalysis. Four articles from the journal *Psychology and Aging* were also excluded for various reasons. One article was excluded because the wrong data set has been shared. Another one was not considered because the provided raw data was only temporarily available. And two other articles were excluded because essential data was missing. The remaining 70 articles were, in principle, eligible for reanalysis.

Most scientific articles contain several statistical inferences, but they are not all equally important. In this study, we focused on statistical inferences that support claims of primary interest. We defined a primary claim (PC) as an a priori hypothesis that is mentioned in the article's Abstract and whose plausibility is being evaluated using null hypothesis significance testing (NHST). Consequently, each PC is underlied by a numerical triplet consisting of test statistic, degrees of freedom, and the corresponding *p* value (e.g., $F(2, 30) = 4.35$, $p = .022$).

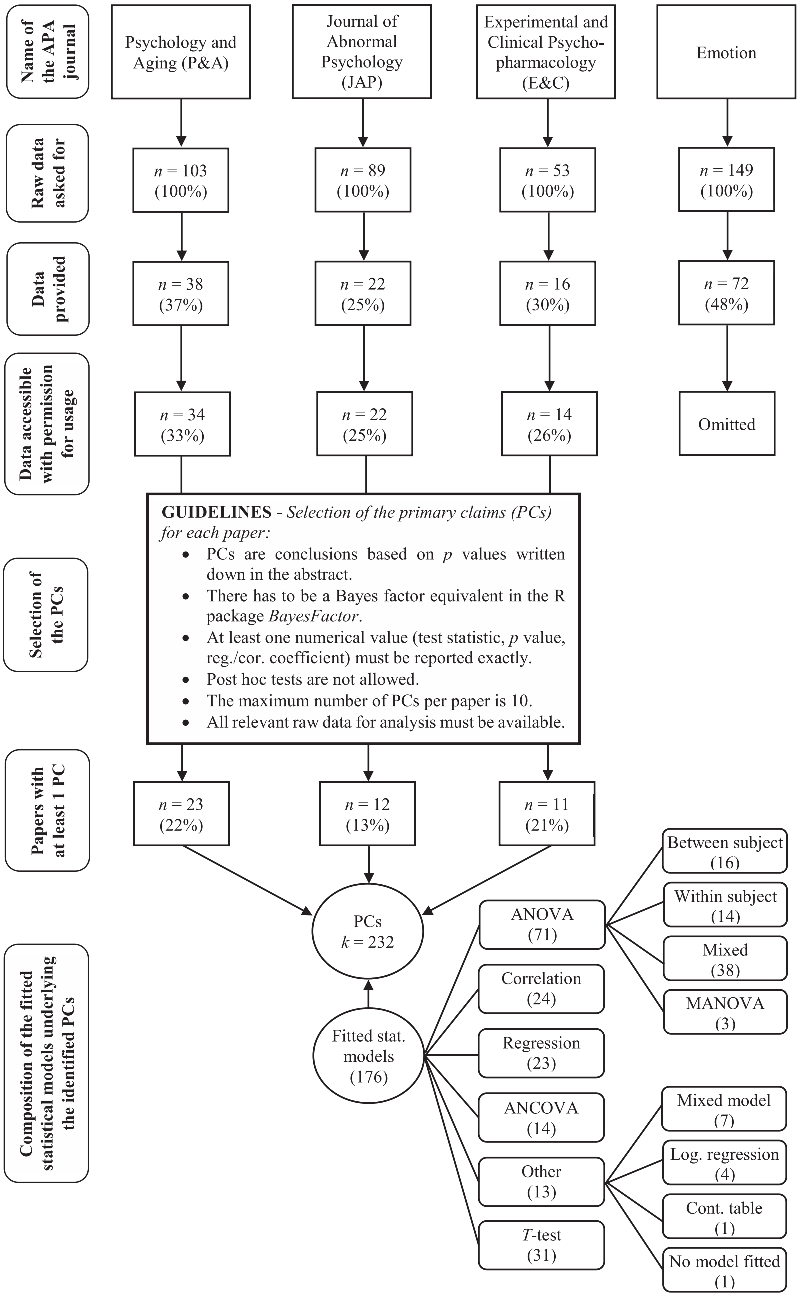
Due to the original goal of comparing the Bayesian and frequentist conclusions, we applied several exclusion criteria to all PCs in our sample of articles. First, we focused only on PCs that were based on models for which a Bayes factor could be computed via the R package *BayesFactor* (Morey & Rouder, 2015, Version 0.9.12–2). The featured models in this package constitute mostly univariate linear models. Table 2 shows the definitions of the (overlapping) type of models for which the BayesFactor package provides support, and the respective R functions we used for reproduction.[5] Second, we only included PCs for which at least one precise numerical value (either *p* value, test statistic, or some coefficient on which the test statistic depends, such as a regression coefficient) was reported. This requirement is a basic prerequisite to reasonably accurately assess reproducibility. Third, we discarded PCs that were based on post hoc tests (because in the light of our original frequentist-Bayesian comparison, it was not clear how to deal with these tests). Fourth, we excluded PCs for which the raw data necessary for reproduction was not available (some authors shared only incomplete data). Lastly, the maximum num-

[4] It is unclear if this article has been peer reviewed. Multiple versions of this article exist and it is listed as forthcoming in the journal *Critical Finance Review* ever since 2017. Here is a link to a version that includes an Appendix with details about the reanalyzed articles: https://pdfs.semanticscholar.org/5c79/5fc4910264e9a9e8dc559e7689594e0f146c.pdf.

[5] Binomial regression was included because, at the start of this study, we speculated that binomial (i.e., logistic and probit) regression models would be included in future versions of this package.

Figure 1

The Count of Articles (n) at Various Steps of the Article Selection Process, the Count of Primary Claims (k) That Were Identified for This Reproducibility Study, and the Counts of the Different Types of Fitted Statistical Models Underlying Them



GUIDELINES - *Selection of the primary claims (PCs) for each paper:*

- PCs are conclusions based on *p* values written down in the abstract.
- There has to be a Bayes factor equivalent in the R package *BayesFactor*.
- At least one numerical value (test statistic, *p* value, reg./cor. coefficient) must be reported exactly.
- Post hoc tests are not allowed.
- The maximum number of PCs per paper is 10.
- All relevant raw data for analysis must be available.

ber of PCs per article was restricted to 10, aiming to strike a fair balance between ensuring that most relevant statistical conclusions are investigated and not skewing our Bayesian comparison too much toward those few articles with an enormous number of eligible statistical results. In the case that more than 10 statistical results fulfilled all criteria of our guidelines, we followed the subsequent Points 1 to 3 consecutively until no more than 10 PCs remained:

1. We select the most important result(s) based on the Abstract and the Discussion section.

2. We focus on the objective/hypothesis that is mentioned first.

3. We select those results that are explicitly mentioned in the text of the Results section.

Table 2

Statistical Model Families for Which Bayes Factors Can Be Calculated via the BayesFactor R Package (Morey & Rouder, 2015, Version 0.9.12–2), Classified by Means of Variable Types (Cont. = Continuous; Cat. = Categorical; Dich. = Dichotomous), Together With the Respective R Functions Used for Reproduction

Type of model	Type of DV	Number; Type(s) of IV(s); additions	R function used
linear correlation	cont.	1; cont.	corr.test()
t-test	cont.	1; dich.	t.test()
regression	cont.	≥2; ≥1 IV cont.; focus on cont. IV	lm()
ANCOVA	cont.	≥2; ≥1 cont. & ≥1 cat.; focus on cat. IV	lm()
ANOVA between	cont.	≥2; cat.; no IV's repeatedly measured	ezANOVA()
ANOVA within	cont.	≥2; cat.; all IV's repeatedly measured	ezANOVA()
ANOVA mixed	cont.	≥2; cat.; some IV's repeatedly measured	ezANOVA()
linear mixed model	cont.	≥1; cont. & cat.; random effects	lmer()
contingency table	dich.	≥1; cat.	chisq.test()
binomial regression*	dich.	≥1; cont. & cat.; logit/probit link function for DV	glm()

Note. DV = dependent variable; IV = independent variable.
* Not included in the BayesFactor package (Morey & Rouder, 2015, Version 0.9.12–2).

All PCs were identified according to these guidelines (see Figure 1 and Table 2) by SS, under the supervision of FTu and WV. After applying these exclusion criteria, we ended up with 232 PCs from 46 articles. Most of the excluded articles used statistical models not included in Table 2, such as structural equation models (SEM), factor analysis, receiver operating characteristic (ROC) curve analysis, latent growth model, and multivariate analysis of variance (MANOVA). In total, six PCs in three articles had to be excluded because the relevant raw data was missing.[6]

Figure 1 shows the composition of the 176 models underlying the 232 PCs. Each of these 176 underlying models was fit to a unique set of data vectors by the respective R function (see Table 2) in light of the reproductions. All 232 PCs rely on simple statistical hypotheses tests (i.e., using a pivotal test statistic, such as the *t*-statistic, whose distribution under H_0 depends only on the sample size) for either one model parameter (193 PCs with $H_0 : \theta = 0$ *vs.* $H_1 : \theta \neq 0$ and 2 PCs with $H_0 : \theta \leq 0$ *vs.* $H_1 : \theta > 0$) or for multiple model parameters simultaneously (37 PCs with $H_0 : \theta_i = \theta_j \; \forall \, i, j \in \{1, 2, \ldots, a\}, i \neq j$ vs. $H_1 : \theta_i \neq \theta_j \; \exists \, i, j \in \{1, 2, \ldots, a\}$ for some $a \in \mathbb{N}$, $a \geq 2$). *t* tests and correlation hypothesis tests always underlie precisely one PC. On the other hand, some of the fitted ANOVAs, ANCOVAs, and regressions underlie multiple PCs. For instance, if both main effects plus the interaction effect of a two-way crossed ANOVA were identified as PCs, then this ANOVA model underlies these three PCs. The 46 articles where at least one PC was identified relied heavily on *p* values and rarely reported effect size estimates. All 46 articles featured purely experimental study designs with human participants. In many cases, repeated measurements were taken on the same variable(s) for each subject, as reflected by the large number of within-subject and mixed ANOVAs.

As is shown in Figure 1, we included PCs for which we ended up fitting MANOVAs (3x repeated-measurement design in two different articles contributing to four PCs) even though our guidelines did not include these types of models (see Table 2). The reason is that during the identification of the PCs, it was not clear that the authors fitted MANOVAs. This only became apparent at the stage of reproduction because the method/result sections were either too vague (Article A) or incorrect (Article B).[7]

Reproduction Procedure

Four members of our team (RA, SG, SS, and TV) worked systematically on the reproductions of the 232 PCs that were identified in the selected 46 articles, using the statistical software R (R Core Team, 2018), often taking advantage of relevant preprocessing work done by FTr. Each article was analyzed according to the following procedure with a "pilot" (lead and first analyzer) and "copilot" (double-check and second analyzer):

1. The pilot of the article read in the relevant raw data and started to reproduce the identified PCs by carefully following the method section and by incorporating all relevant information in other parts of the article such as the Introduction, the Results, and Discussion sections, and footnotes. If necessary, information in referenced articles was incorporated too. The complete code of analysis was saved as an R file. In addition, pilots produced a brief report for each article, detailing the main findings of their reproduction attempt.

2. The article was transferred to the copilot, who continued to work on the reproductions of the PCs. If the pilot was able to reproduce a PC, the copilot checked the analysis

[6] We did not formally study the interrater reliability in selecting the article's PCs based on our guidelines, despite the fact that linking conclusions in the abstract with statistical results in the body of the article was not always straightforward. However, when performing the reanalyses, the analysts assessed the correctness of the application of the guidelines for each article and found no issues with the selection of PCs. The only exception was the removal of the six PCs for which the relevant data had not been provided, a fact of which SS was not aware while selecting the PCs.

[7] Three of these four PCs (2x Article A, 1x Article B) have exactly the same *F*-test statistic when a mixed ANOVA is fitted instead because they are either associated with a between-subject factor (2x) or with an interaction of a between-subject factor and a within-subject factor with only two levels (see, e.g., Maxwell et al., 2018). We labeled all three models as MANOVAs because the authors of the two articles always chose the multivariate *F*-ratios in case they differed to the univariate ones (i.e., for testing within-subject factors with more than two levels).

code for errors and plausibility. In the case of nonmatching results, the copilot tried to achieve reproduction by writing new R code into a separate file. The potentially different conclusions and the summary of this second attempt were then added to the report. Never did the copilot alter the pilot's code or report, not even in case of mistakes.

3. In the case that neither the pilot nor the copilot was confident that their code of analysis exactly corresponded to the descriptions in the article or was error-free, a third team member went through their code of analysis and, when necessary, wrote additional R code.

4. Eventually, if, despite all efforts, a PC could not be reproduced, the team members tried to identify (plausible) reasons for the irreproducibility.

Assessing the reproducibility of the 232 PCs turned out to be a challenging and laborious endeavor, and we went great lengths to arrive at successful reproductions for all 46 analyzed articles. When an initial reproduction attempt failed, we employed the following strategy. First, we deviated from the article's description to get matching numerical results and to pinpoint the reasons for nonmatching reproductions. In some cases, statistical models that were close but not identical to the ones reported in the article were fitted, such as models with additional or fewer covariates/interactions. In other cases, various plausible subsets of participants (e.g., not applying a reported exclusion criterion or removing cases with missing values in some variables) were used in case of nonmatching degrees of freedom. Second, we used every reported numerical information regarding the used variables to pinpoint the reason(s) for deviating results. In particular, we tried to reproduce numerical values reported in the article that were not the prime target of reproduction but seemed relevant for this PC (e.g., means, standard deviations, sample sizes, demographic information about the subjects) as doing so could help in figuring out what the authors calculated exactly (e.g., checking the reported descriptives might help to identify which participants were included in the analysis). In cases where little or no additional numerical information besides test statistic, degrees of freedom, and p value was reported, we tried to reproduce relevant figures from the article to assess whether we were working with the same raw data or not. These additional analyses were neither done exhaustively nor systematically. We did not contact the authors of the article in case of reproducibility issues. The reason is that we believe that, ideally, reproducibility should be possible without the assistance of the authors, and this study assesses the extent to which reality deviates from this ideal situation.

Once all 46 articles were analyzed by a pilot, a copilot, and if needed, a third pair of eyes, RA and TV summarized the reproducibility results. The systematic summary contains the following information about each PC:

- Model information: the type of the statistical model and the type of variables it includes, the effect of interest (main or interaction effect?), the distribution of test statistic under H_0;
- Numerical information: reported and reproduced values of test statistic, degrees of freedom, and p value;
- Verbal descriptions: information about potential mismatches between reported and actual calculation, issues encountered during the reproduction attempt, numerical checks of related and underlying numerical results, suspected reasons for irreproducibility, and additional information about the PC.
- As a measure of numerical precision, the absolute (value of the) relative deviation (ARD) between original and reproduced result was then calculated for each PC as follows:

$$\mathrm{ARD}: = \left| \frac{T_{\mathrm{reproduced}} - T_{\mathrm{reported}}}{T_{\mathrm{reported}}} \right|, \qquad (1)$$

If a test statistic was reported in the article, T_{reported} represents its value and $T_{\mathrm{reproduced}}$ represents the value of the test statistic in the reproduction rounded to the same number of digits as $T_{\mathrm{reported.}}$ If the value of the test statistic was not reported, ARDs were calculated based on regression coefficients (35×), standardized regression coefficients (4×), odds ratios (1×), correlation coefficients (24×), and, as a last resort, p values (2×), again using Equation 1 and again rounding the recomputed values to the same number of digits as in the reported values. The p value was not used as the primary number to quantify the numerical precision via Equation 1 because it was either not reported precisely (155 PCs) or reported with fewer or equal significant digits[8] as the test statistic or the reported coefficient on which the test statistic depends (74 PCs).

Further, we checked whether the reproduced degrees of freedom match with the reported ones. This was impossible for 87 of the 232 PCs, most of them coming from regressions and correlation tests, for which no degrees of freedom were reported. Not reporting degrees of freedom was quite common for t-statistics (e.g., a linear regression result of $\beta = -4.543$, $SE = 2.12$, $t = 2.143$, $p < .05$).[9]

Quantifying the Magnitude of Deviation

Some authors (e.g., Hardwicke et al., 2018) use the ARD to classify discrepancies as minor or major numerical errors depending on whether or not the obtained ARD exceeds a chosen threshold for a variety of quantities such as sample sizes, standard errors, test statistics, effect size estimates, and p values (e.g., Hardwicke et al., 2018 used a single cutoff of 10% for these widely different scales). However, ARDs computed on different scales (such as F and t statistics, regression coefficients, correlation coefficients, and p values) are not commensurable. This is easily demonstrated with the following fictitious example where the article reports $t(20) = 2$ for a PC, and

[8] For example, the significant digits of the test statistic and the p value in $t(30) = 2.452$, $p = 0.02$ are 4 and 2, respectively.

[9] Although the degrees of freedom of a t-statistic are a function of the sample size (e.g., $df = n - 2$ in the case of a correlation test), we did not compare the sample size reported in the article with ours as a substitute in case no degrees of freedom but a sample size was reported for a PC. Such a comparison would be problematic because it is often not clear whether the reported sample size represents all cases or only those cases without missing values or specific characteristics (e.g., that some variable lies within a certain range). In linear regression, for instance, all cases with at least one missing value for any of the used variables are dropped, and the actual sample size represents all cases without any missing values.

the reproduction produces $t(20) = 1.8$ yielding an ARD of 10%. Because $\sqrt{F(1, n)}$ follows the same distribution as $|t(n)|$, the article could have chosen to report $F(1, 20) = 4$ instead. If then the reproduced value for $F(1, 20)$ is 3.24 ($= 1.8^2$), this results in an ARD of 19%. Equally, the authors could have chosen to report the (two-sided) *p* value of 0.059 instead. The ARD would then be approximately 47% because the reproduced rounded *p* value is 0.087 (based on $F(1, 20) = 3.24$). Because of this lack of commensurability, we do not classify the PCs into ordered categories of severity of deviations via the ARD.

To create commensurable quantifications of the magnitude of deviation via some common distance measure such as the ARD, the (to be) reproduced values need to lie on a common or at least on a similar scale for each PC. Suitable common scales for NHST results are effect size scales. Rosnow and Rosenthal (2003) classify effect size scales into three families. The *d* family which includes standardized scales such as Cohen's *d* and root mean squared error standardized effect, the ρ family which includes scales that measures the strength of association such as (partial) η^2, and (partial) ω^2, and the ratio family which consists of the relative risk and the odds ratio. Members of the ρ family are particularly suited to be used to quantify the magnitude of deviation (e.g., via the absolute deviation) because they are bounded between zero and one. Unfortunately, the reported information was not sufficient to calculate effect size measures for many PCs (e.g., due to missing degrees of freedom), making it impossible to transform all PCs onto a common scale.

A second, more obvious, common scale to all analyzed PCs is the *p* value scale. Hardwicke et al. (2018) assessed for each selected hypothesis test whether or not the reproduced *p* value falls on the other side of the 0.05 threshold compared with the reported *p* value. Because, in the 46 analyzed articles, *p* values were solely computed to assess whether or not they passed the 0.05 threshold, we will adopt this dichotomization of irreproducibility. As such, when we encounter a deviation between the reported and the recomputed result, as indicated by an ARD above zero, we will classify a change in significance as a *decision error*.

Transparency

The study of Vanpaemel et al. (2015) received approval from the Ethics Committee of the Faculty of Psychology and Educational Sciences of the University of Leuven under the restriction that it will not be disclosed who shared their data and who did not. Only those researchers working directly with the data gathered by Vanpaemel et al. (2015) know for which of the 394 articles raw data was shared. In order to not constitute a breach of confidentiality, we do not reveal which articles were reanalyzed. We regret that this implies our study is, in itself, not reproducible, as it is impossible for others to check the correctness of our reanalyses. We have tried to accommodate this by providing concrete (anonymized) examples of the problems we encountered below. If individual authors whose article might be included in our study are interested in the reproducibility status of their article, we invite them to contact us, so that we can share the code and conclusions of our reanalysis of their article with them.

Results

Summary of the Reproductive Success

Our main results are summarized in Figure 2. As shown by uncolored cells in Figure 2, we were able to successfully reproduce 163 (70%) of the 232 PCs, in the sense that the ARD is zero. However, 26 of these 163 PCs do not qualify as successful reproductions in the strictest sense. For seven PCs, marked with an X, the ARD is zero, but we could not reproduce the reported degrees of freedom. The reproductions for four of those seven PCs produced the reported *p* value, but the reported test statistic and degrees of freedom are inconsistent with the reported *p* value, implying that the reported degrees of freedom are typos. For the three other PCs, no *p* values were reported, making a complete consistency check impossible. For two of these three PCs, we believe that the reported degrees of freedom are typos in the sense that they were not used by the authors to compute the *p* value. This belief is based on the fact that they are equivalent to the reported degrees of freedom of another PC from the same article, which suggests a copy-paste error.[10]

For another 18 PCs, marked with an exclamation mark in Figure 2, the ARD of zero could only be achieved by deviating from the article's method description. For 13 PCs we needed to deviate from the reported exclusion criteria to achieve reproduction, for two PCs we were successful by using the binned version of a variable instead of its raw version (unlike what was reported in the article), for one PC we had to use the Greenhouse-Geisser correction (unlike what was reported in the article), for one PC we only had to use cases where a factor variable had a certain level and not all cases (as was implied by the method description), and for one PC we had to use a multivariate test statistic instead of the reported univariate test statistic.

Finally, for one PC, marked with an "S," the ARD of zero was only achieved by taking the absolute value of the reported *t*-statistic. We chose this because, for one, we identified the sign of the reported *t*-statistic as a typo, because the reported corresponding regression coefficient had the opposite sign and, second, a two-sided test was performed which meant that the *p* value did not depend on the sign of the *t*-statistic.

As indicated by colored cells in Figure 2, we were unable to achieve an ARD of zero for 69 (30%) of the 232 PCs. For one PC, the variable and data selection, and also the description of the statistical model fitted, was so unclear and internally inconsistent that we refrained from fitting and reporting the result of a specific model. For three of these 69 PCs, marked with the exclamation mark, we used a sample that violated some of the reported exclusion criteria, as this resulted in matching degrees of freedom and other PCs from these two articles were successfully reproduced by using that sample.[11]

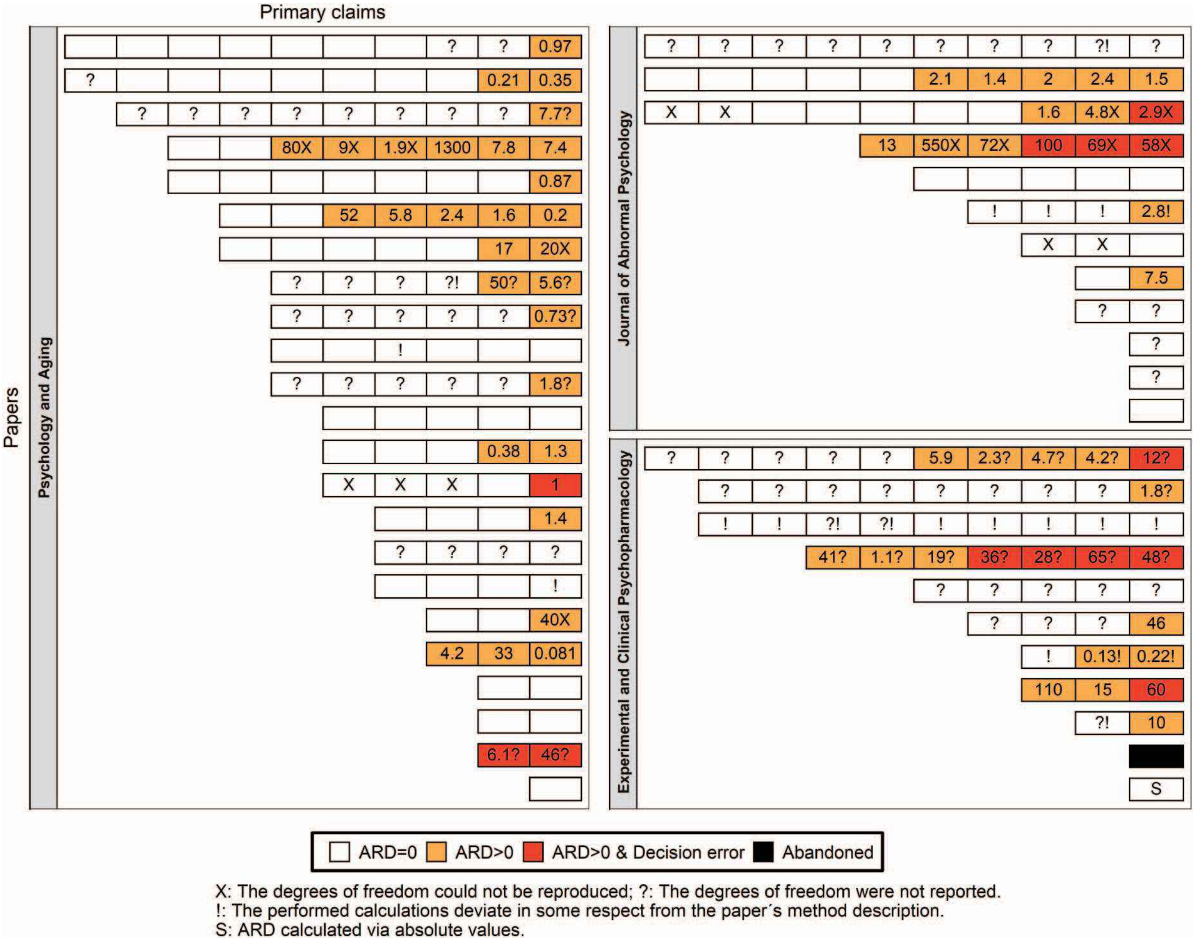
For 185 of the 232 identified PCs (including the abandoned PC), a *p* value of less than or equal to 0.05 was reported in the article

[10] It is possible that two test statistics are very similar (i.e., equal when sufficiently rounded) even though a different subset of participants was used for their calculation leaving us only with the certainty that we are either dealing with an inconsequential incorrect calculation or a typo.

[11] Technically speaking, these two articles contain contradictory information because their exclusion criteria did not match the reported degrees of freedom, which are a function of the sample size. Hence, some information in the article is violated, regardless of which sample is used for reproduction.

Figure 2

Absolute Relative Deviation (%) of the 232 Primary Claims That We Identified in 46 Articles Published 2012 in Three APA Journals: Psychology and Aging (P&A), Journal of Abnormal Psychology (JAP), and Experimental and Clinical Psychopharmacology (E&C)



X: The degrees of freedom could not be reproduced; ?: The degrees of freedom were not reported.
!: The performed calculations deviate in some respect from the paper´s method description.
S: ARD calculated via absolute values.

Note. ARD = relative deviation. See the online article for the color version of this figure.

and interpreted as statistically significant (one PC had a p value of exactly 0.05). Excluding the abandoned PC, we encountered 13 decision errors, all of which with reproduced p values larger than 0.05. This means that 7% of all decisions to reject the tested null hypothesis could not be reproduced. Conversely, none of the 47 PCs with reported p values larger than 0.05 achieved significance in our reproductions.

We estimate the total amount of time spent on the reproductions of the 232 identified PCs to be 280 workdays, vastly exceeding our worst predictions. This enormous duration is not just the result of the 69 PCs we were unable to reproduce. The 163 reproduced PCs contributed considerably too. Only for a tiny minority of them was the successful reproduction as straightforward as reading in the data and exactly following the analytical steps as detailed in the article. In contrast, most reproduced PCs are the result of the laborious task of reverse-engineering the data analytic steps, an effort that took multiple days if not weeks of work. Successful reproduction of a PC was often only achieved by the copilot, and for a handful of PCs, only the third pair of eyes was able to reconstruct the calculations conducted by the authors.

In what follows, we discuss the likely reasons for the 69 reproductive failures as well as how vagueness made 163 reproductive successes hard gained.

Reasons for Reproductive Failure and Identified Errors

For some of the nonreproduced PCs, we could pinpoint errors in the statistical analysis or the reporting of the results as the reason(s) of irreproducibility. In light of our reanalysis, we also identified many incorrect statistical results in the articles that were not part of the 232 PCs, such as summary statistics (e.g., means, standard deviations, sample sizes) and other NHST results. Further, we sometimes were able to identify incorrect verbal descriptions of the conducted statistical analysis, as indicated by the exclamation mark in Figure 2. Here, we present a list of error types that contain many of the identified errors:

1. **Rounding of numerical output that had already been rounded.** It is well-known that repeated rounding can

lead to an accumulation of numerical errors. How repeated rounding can lead to an incorrect result becomes clear in the following example: If the test statistic is $T = 3.41461880\ldots$, it is shown in SPSS as a rounded result of $T = 3.415$. Taking the SPSS output and rounding it once more before reporting leads to the erroneous result of $T = 3.42$ in the article.

2. **Using rounded values in calculations.** This problem is related to the first one. In subsequent steps of a calculation, the most accurate representation of a number should be used. However, some authors used rounded values for the calculation of means and standard deviations or to fit a statistical model. In some cases, variables were constructed based on the rounded values of other variables.

3. **Inconsistency in the selection of cases or variables.** The selection of data (cases or variables) for which the authors fitted the statistical model is inconsistent with what has been described in the article. We encountered incorrectly applied filter variables, incorrect selections of subsets, and wrongly selected variables. In one case, the used sample size was inflated because it included empty rows in the data file. Obviously, in most of those cases, we are not able to determine whether the authors erred in their data analyses or in the reporting of their methods. As mentioned above, 13 PCs (coming from three different articles) had a mismatch between the participants included in the analysis and the exclusion criteria reported in the article, which were of the type "Only participants with a score higher than X in variable Y are going to be used for the analysis."

4. **Incorrect labeling of variables or numerical results.** Examples are the switching of column labels inside a table, the switching of the labels of two experiments, incorrect information in the caption of tables or figures, and the mixing up of standard deviation (*SD*) and standard error (*SE*) which can easily go undetected for small sample sizes. We succeeded in finding these errors by first spotting implausible values and then finding related variables/results with precisely that value.

5. **Typos.** Typos are estimated to occur in about 0.2% of the words in written scholarly texts (see Pollock & Zamora, 1983) and we do expect similar error rates for numerical results. Reported results that are simply impossible (e.g., a p value that is larger than 1 or negative) must be typos. However, most of the time, only strong indications but not certainties exist regarding deviating results. Examples are transposition errors (reporting 0.045 as opposed to 0.054) and transcription errors (reporting 2.742 as opposed to 2.7042) within the range of possible values.

6. **Copy-paste errors.** Copy-paste is convenient for typewriting, but it is also a source of mistakes. An incorrect statistical result can be copy-pasted (e.g., an unintentional mixing up of MANOVA and ANOVA F-ratios or of error degrees of freedom from different within-subject factors in the SPSS GLM output). We suspect that authors frequently copy-paste and adapt the statistical output of a neighboring result out of convenience. However, if the result is subsequently not correctly altered, an error is committed. One example would be to forget to replace the degrees of freedom when copy-pasting the main effect $F(2, 32) = 5.3$, $p < .05$ of an ANOVA to faster report on the interaction effect. Copy-paste errors were identified and discussed by Bakker and Wicherts (2011) who conclude that they are quite common.

Besides errors from the above six categories, we could sometimes identify a reported statistical result as incorrect despite not knowing the reason for the deviation because it contradicted other related results, such as a summary statistic of the underlying raw data or another statistical result from the same model. Our reproductions revealed a surprisingly high number of glaring reporting errors, for which no raw data is needed to identify them as mistakes. Examples are inconsistent test statistic, degrees of freedom, and p value triplets, numbers that are outside the possible range of values (e.g., positive values for the logarithms of numbers that are smaller than one), opposite signs for regression coefficients and the corresponding t-statistic, and the reporting of non-matching numbers for the exact same statistical result in different parts of the article. Some of these glaring errors even made it into the abstract of the article.

Pinpointing the reasons for deviating results was impossible for many PCs. The reason is that the origin of a PC was often obscured due to vague data, variable, and model descriptions, an issue that is discussed in-depth in the next section. Software differences (we used R, whereas the authors in the reanalyzed articles used SPSS, SAS, Excel, Mplus, and GraphPad PRISM) are unlikely to explain encountered differences in numerical values. What differs between various statistical software when it comes to t tests, ANOVAs, and regressions are the default settings (e.g., TYPE I vs. TYPE III ANOVA or maximum likelihood vs. unbiased variance estimator in regressions), and we always computed all alternative model specifications if the initial reproduction attempt failed.

Vagueness Makes Assessing Reproducibility a Nightmare

While a 70% reproduction rate might seem reason for light optimism, we fear such optimism is not warranted. The reason is that most successful reproductions are predominantly the result of tedious and time-consuming work. This was due to the fact that that information about the provided raw data was often difficult to understand, and information about the relevant variables, data manipulations, and the used statistical model was often vague or inaccurate. Lacking a culture of data sharing and reanalysis in psychology, it is fair to assume that most of the corresponding authors that shared their raw data with Vanpaemel et al. (2015) did not take special efforts to make the shared data easy to use and understand for independent reproducers. We believe, however, that even if all analyzed data sets in this study had been well-structured and well-documented, the reanalyses would not have taken considerably less time.

To accurately repeat the calculations underlying the PC of interest, it is necessary to know the used variables and the used cases of each variable, all relevant transformation and missing-data imputations of these variables, the nature of the used statistical model, and the used estimators. These four categories will be defined and elaborated below with some examples.[12]

1. Information about the **provided raw data**. For each relevant variable, it must be known what the corresponding raw data vector is and what its entries represent.

2. Information about the exact **handling of the provided raw data**. In situations where the included variables are derived from other primary variables, information about the necessary calculation(s) must be provided (e.g., the false positive rate is calculated by taking the ratio of all successfully identified cases and all identified cases for each participant). Ideally, this information is provided even for seemingly obvious calculations, as there might be differences in tacit assumptions between researchers and scientific fields. If only a subset of cases is used for the calculation, it must be clear how this subset is composed (e.g., only participants older than x years, only measurements in Condition A). Furthermore, it must be clear how to deal with missing values (e.g., case-wise deletion).

3. Information about the exact **statistical model**, the dependent and independent **variables** it includes and the variable(s) associated with this **hypothesis test**.

4. Information about the calculation of the test statistic and the estimation of the parameters determining its distribution under H_0. This includes information about how standard errors and the degrees of freedom are estimated. The specific types of **estimators** used should be mentioned, and ideally, a justification for that choice, particularly in the case of nonstandard estimators, should be given.

Let us start with Category 1 (provided raw data). For 25 of the 46 articles, the raw data was provided via SPSS (.sav). For another 18 articles, it was provided in the form of an MS Excel file (.xls,.xlsx) or in comma-separated value format (.csv). For the remaining three articles, the raw data was provided as.txt files. The majority of provided raw data sets had a comprehensible structure and self-explanatory column names. However, around one third of all provided data sets contained either cryptic column names or cryptic factor levels or both. This complicated assessing reproducibility considerably, especially if the dataset included many variables (i.e., columns). Column names such as “ckettrm1,” “mck3,” “PCLR-03,” and “OxiCtPl” and factor levels such as YK, YM, AK, AM are ambiguous and should come in conjunction with a detailed description of their meaning. Numerical factor levels such as 1, 2, 3, 4 are even more mysterious. For instance, some of the provided data sets included dummy variables where it was not clear what the zeros and ones stood for. Databooks or metadata with details about abbreviated variable names and factor levels have not been provided for many data sets in this sample of 46 articles, and if so, the information was often not sufficiently detailed. We managed to reproduce PCs with unclear variables and factor levels mainly by looking at reported summary statistics and by fitting multiple plausible models.

A common Category 2 (handling raw data) issue was finding out which cases were used in the original analysis. If the reported degrees of freedom did not match with the use of all cases contained in the dataset and no further (or incorrect) information was given, we tried to get matching results by using reasonable subsets. However, testing all possible subsets was not always a feasible strategy. If there was no indication of which cases had been used, and the number of cases was large and significantly greater than the specified degrees of freedom, the number of meaningful subsets was simply too large. Reasons for ambiguity can be imprecise descriptions on the handling of outliers. One article, for example, stated that outliers identified using a fixed z-score as a cutoff were removed. It was, however, not clear whether this was done for all or only for some of the relevant variables.

Once having identified the used cases, it has then to be assessed how the provided numerical information maps to the variables used in the statistical analysis. The following two cases are illustrative examples of missing or incomplete information on data handling:

- One of the provided data sets included a participant with missing values for two relevant variables, each of them part of a PC. The article did not mention this participant at all. It turned out that one PC was calculated by fitting a model without this participant. In case of the other PC, we found out by trial and error that the relevant missing variable (a sum of eight dummy variables also included in the provided data set) was imputed by summing only those variables (six out of the eight) that did not contain missing values for this subject.
- Accuracy (d') and response bias (c) had to be calculated based on the hit rate (HR) and the false alarm rate (FAR). The article only references a book with the standard definition of $d' = z(\mathrm{HR}) - z(\mathrm{FAR})$ and $c = -0.5 \cdot (z(\mathrm{HR}) - z(\mathrm{FAR}))$ where $z(\cdot)$ is the quantile function of the standard normal distribution. However, d' could not be calculated like this with the provided data set because it included extreme values for HR and FAR (i.e., 0 or 1), resulting in a nonfinite quantile function. A literature research on d' and c revealed multiple ways to calculate d' and c. Trial and error revealed that the authors used the “loglinear” approach (described, e.g., in the unreferenced Stanislaw & Todorov, 1999, p. 144).

Overall, figuring out the necessary Category 3 information (statistical model, hypotheses, and tests) was the most time-consuming task. Essential information about the statistical model was either missing or hard to extract for the majority of analyzed articles because descriptions of manipulated and observed variables were mostly presented verbally, without equations or symbolic notation for the important variables and models. Furthermore, many articles described a handful of fitted models in one part of the article (often the Method section) and a myriad of test statistics and p values in another part (mainly the Results section),

[12] Whenever a article is quoted, the quote is rephrased in order not to reveal the identity of the article and the authors.

without a clear indication of which result belongs to which model and which variable. The mapping between models and the results of hypothesis tests was especially hard to decipher when only the results of some of the fitted models were reported. The following article, for example, contains four hypothesis tests, but only two test statistics are reported: “We observed a significant main effect of sleep condition for accuracy and frequency for Group 1 and Group 4 (F(1, 12) equals 6.4 and 5.8, respectively).” In the case of ANOVAs, the description was often not clear about whether interaction effects were included or not. If in case of unclarity, no interaction effect is being reported in the result section, it might not have been included in the model, or it might have turned out to be insignificant and deemed unworthy of a report. Only a small number of articles made reproduction easy by reporting fitted models in a formal language (e.g., *Model* III: $Y \sim X_1 + X_2 + X_1 {}^* X_2$ in the Method section and *Model* III found an interaction effect of X_1 and X_2, $F(1,42) = 6.23$, $p < .05$ in the Results section).

One source of confusion was the use of adjectives that are ambiguous in a statistical context such as “hierarchical,” “mixed,” and “random.” Some authors referred to the variable selection procedure of nested models as hierarchical models, whereas others used hierarchical to signal that some variables are nested within other variables. The modifiers mixed and random were used to describe linear mixed models by some authors and repeated measures ANOVAs by others.

What follows are two illustrative examples of unclear model descriptions:

- One article states that “Regardless of wealth, subjects rated their own drawings less attractive than those created by a friend” and our reproduction showed that the authors had fitted a model that does not include the variable wealth. However, one could also interpret this statement as a description of a significant main effect for the factor (oneself vs. friend), despite wealth being included in the model, or no interaction effect between wealth and this factor, or both.
- In another article, the authors stated: “The product terms were residualized relative to lower-order terms to facilitate the interpretation of main effects in each model and to avoid multicollinearity problems.” Based on such a verbal description in the Method section, it is very hard to figure out what the underlying statistical model is. The reproduction of the PC related to this model description was achieved by fitting a classical linear regression. Another reported statistical result related to this model (not identified as a PC), on the other hand, could not be reproduced. Hence, we are still uncertain about the meaning of this description and the exact model they fitted.

A final necessary piece of information needed to solve the reproduction puzzle involves details regarding the calculation of the test statistic and the estimation of the degrees of freedom (i.e., Category 4 information). Commonly used estimators for degrees of freedom are the Welch-Satterthwaite approximation for t tests, the Greenhouse-Geisser and Hyunh-Feldt corrections for ANOVAs, and the Satterthwaite’ and Kenward-Roger’s approximations for linear mixed models. These estimators typically produce non-integer degrees of freedom, indicating their use without an explicit statement. What made reproductions sometimes delicate is that precise non-integer estimates for the degrees of freedom were rounded to and reported as integer values either by hand or by software. If the used correction is not mentioned in the article, the reader might then falsely assume that degrees of freedom are computed without any correction, meaning that they are just a function of the sample size and the levels of the respective factor and conclude that not all cases had been used.

Regarding Category 4, information about the precise estimators used was rarely given in the analyzed articles. Almost none of the analyzed articles mentioned whether Type I, II, or III sums of squares were used for ANOVAs in case of unequal group sizes. The used residual variance estimator in regression analysis was never reported (see below). Details about estimators in linear mixed models were scarce (e.g., REML vs. ML). We observed that the vast majority of statistical analysis we were able to reproduce were conducted by using the default settings of the respective statistical software (e.g., TYPE III ANOVA in SPSS). For this reason, we assume that at least some of the authors were not aware of default settings and potential statistical alternatives. This is unfortunate because the default choices made in the software packages can have a significant impact. For example, Mplus (Version 6) computes as a default the maximum likelihood estimator (assuming a Gaussian likelihood) to estimate the variance of the residuals in regressions, which is a biased estimator of the population variance. This estimator differs from the more commonly used unbiased estimator by a factor of $\frac{n-k}{n}$. In the case of a low number of participants n and a high number of estimated regression coefficients k, this difference is non-negligible and it can cause p values to differ and potentially become smaller or larger than .05 depending on the used estimator.

Discussion

The vast majority of successful reproductions in this study are the result of a painful and frustrating process of trial and error, because of the existence of multiple, plausible data analytic pathways to calculate the numerical triplet of the primary claim that are compatible with the provided raw data combined with the vague description of the statistical methods in the article. Ideally, published statistical results are both correct and easily verifiable by another reasonably skilled researcher. Only when sufficient details are available can a reproduction attempt conclude whether or not the reported result is correct. In the case of numerous possible ways to calculate a reported result without violating any information regarding its calculation, the correct way might be missed if “only” a few of those ways were tried. What makes matters worse is that if an analysis (out of a large number of attempts) yields a matching result, it cannot be ruled out that the reproduction was due to coincidence, especially when dealing with rounded values.

The lack of precision in the description of the origin of a reported statistical result makes its verification difficult, time-consuming, and frustrating. From this study, we conclude for the APA journals we investigated that reproducing the reported statistics based on the raw data and the described methodology from the article is similar to the reconstruction of the route taken by a walker in the garden of forking paths (Gelman & Loken, 2013).

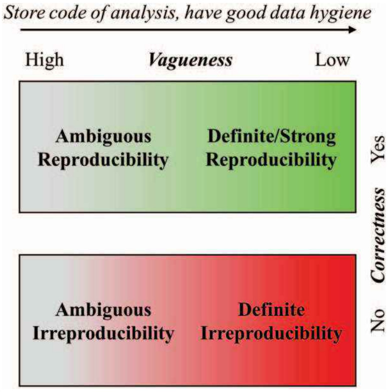
A Two-Dimensional Classification of Reproducibility

To properly address the challenges associated with verifying published statistical results, we propose a novel method for classifying the reproducibility of numerical results (see Figure 3). This classification is done by evaluating two dimensions: vagueness and correctness.

A reported number is **correct** if *it can be reproduced without violating the reported information regarding the underlying calculations*. This dimension is dichotomous if the number of interest is the result of nonstochastic calculations. The result from the calculation 2 + 2 is exactly 4 and every other value, be it 20, 3.9, or, 4.0001, is incorrect. In practice, it will be useful to define a margin of error around the true value (e.g., via the absolute (relative) deviation) and to consider all values within that interval as correct. By doing this, one allows the differentiation between meaningful errors and minuscule deviations stemming from software differences, rounding, or something else.

The degree of **vagueness** is determined by *the amount of information provided about all necessary steps to calculate the reported number from the underlying raw data*. Unlike correctness, this dimension is not dichotomous. A natural way to quantify vagueness is to define a vagueness score that is equal to the precise number of possible ways that are not in disagreement with the available information. A vagueness score of 1, then, signifies the absence of any ambiguity about the origin of the reported number, while a vagueness score of 33 means that there exist exactly 33 different ways to calculate it that do not contradict the provided information. If, for instance, the relevant variables and the corresponding columns in the data set, and also the handling of potential variable transformations and missing values are precisely described for a two-sample *t* test, but no information regarding the assumption of equal variances is given, we are left with only two possible calculations (i.e., the vagueness score equals 2).[13] In practice, researchers will not be eager to spend time counting all possible ways to arrive at a result. Fortunately, this is not necessary as the most practical classification of a reproducibility attempt is achieved by dichotomizing the vagueness dimension into high and low. Two natural dichotomization approaches exist. A strict approach that only classifies cases with absolutely no ambiguity as low, and a more lenient one, where all cases with a small number of easily computable ways fall into the category low.

Figure 3
Classification of Reproducibility via ***Correctness*** *of the Reported Value and the Amount of Information Provided to Repeat the Underlying Calculations (****Vagueness****), Together With Ways to Reduce the Vagueness of the Performed Calculations*


Store code of analysis, have good data hygiene
High
Vagueness
Low
Ambiguous Reproducibility
Definite/Strong Reproducibility
Ambiguous Irreproducibility
Definite Irreproducibility
Yes
Correctness
No

See the online article for the color version of this figure.

Once it has been determined when the reported number of interest is considered as correct or incorrect and when its vagueness is considered as low or high, a reproduction attempt puts it into one of four categories (see Figure 3): (a) *definite/strong reproducibility* (a correct number and low vagueness); (b) *ambiguous reproducibility* (a correct number and high vagueness); (c) *definite irreproducibility* (an incorrect number and low vagueness); and (d) *ambiguous irreproducibility* (an incorrect number and high vagueness).

In the case that a reanalyst is not infinitely skilled and flawless but human, it is obviously possible that mistakes were made in the reproduction attempt of a reported statistical result (e.g., a *p* value, a sample mean, or an estimated effect size) that resulted in a miss-classification. For the sake of simplicity, it is assumed that an error-free reanalysis was conducted for the remainder of this paragraph. In the case of a definitely irreproducible result, it can then be concluded that the reported number is incorrect. On the other hand, verifying the incorrectness of an ambiguously irreproducible result is close to impossible because it cannot properly be differentiated between a reporting error and the failure of the reanalyst to repeat the true underlying calculations. Strongly reproducible results are what we should strive for. If a statistical result is strongly reproducible, the reported number is correct and sufficient transparency about its calculation, allowing its verification by independent researchers within a reasonable amount of time, is available. Ambiguously reproducible results are to be interpreted with great care, especially if the prespecified margin of error is wide. In the case, that one chooses to classify a statistical result without dichotomizing vagueness, it applies that the higher the vagueness, the lesser the difference between reproducibility and irreproducibility (hence the greying out in Figure 3) and the smaller the merit of a reproduction attempt as it is just not clear what exactly to reproduce.

How (Strongly) Reproducible Is Psychology?

Sixty-nine (30%) of 232 identified key statistical claims from 46 articles could not be reproduced exactly, despite a huge investment of our time and effort. For some of these claims, this resulted in a loss of statistical significance. Further, we encountered incorrect method descriptions and errors in the conducted data analyses and the reporting of statistical results.[14]

To interpret our findings we turn to the findings of Hardwicke et al. (2018), the only other investigation of reproducibility via raw data in psychology. Inspection of Hardwicke et al.'s (2018) reanalysis scripts (https://osf.io/q4qy3/) reveals that their sample of articles contains very similar issues than the sample of articles in

[13] In this example, it is assumed that the researcher either used the classical *t*-test or the Welch-test but no other (unconventional) way to estimate the pooled variance of the two groups.

[14] Although it is tempting to summarize the severity of encountered discrepancies in a single number and reproducibility on the article level via some categorization, we will not do so because it would give an unjustifiable impression of precision.

this study, with the rate of problems being a bit higher overall. The identified number of reproducibility issues in their sample of articles is particularly surprising because the authors were required by the journal to share their raw data publically. This indicates that authors do not necessarily go to great lengths in the avoidance of reporting errors just because other researchers can redo and verify their statistical analyses without first having to ask for the underlying raw data. Somewhat surprisingly, they did not encounter a single decision error despite having recalculated 185 *p* values. However, the absence of decision errors in this sample of *p* values might be a statistical fluke considering that 17 of these 185 *p* values had ARDs larger than 10%. Future research might be able to tell if reporting errors are particularly likely to be inconsequential.

Both Hardwicke et al. (2018) and our study made use of the raw data. However, as mentioned in the literature overview (see also Table 1), several studies looked at the consistency of the reported results in the articles. In the sample of the current study, a precise triplet of test statistic, degrees of freedom, and *p* value was reported for 52 (32%) of the 163 successfully reproduced and for 23 (33%) of the 69 PCs that we could not exactly reproduce. Five of these 52 and none of these 23 PCs had inconsistent *p* values. Four of the five PCs are inconsistent because the reported degrees of freedom are incorrect. The fifth PC is inconsistent because the reported *p* value is incorrect. However, the numerical difference is small and inconsequential and can be explained by calculating the *p* value with the rounded test statistic. These observations suggest that there is no evidence that key statistical results with consistent *p* values are more likely to be reproducible than inconsistent ones. In any case, *p* value consistency checks are no substitute for a reanalysis that starts from raw data if one wants to verify the conclusions drawn via NHST.

Generalizability of the Current Study

There are at least two ways in which our current study may fail to generalize broadly. First, the articles reanalyzed may not be representative of the full set of articles published in 2012 in one of the three analyzed journals. Second, the current study may provide limited information on the reproducibility of comparable articles that were published more recently.

The results of our article are likely positively biased due to two selection effects. First, by design, we only assessed the reproducibility of the PCs for which the authors were willing to share their raw data. We suspect that the willingness to share raw data is negatively related to the reproducibility of the results, primarily because authors that are unsure about the correctness of their statistical analyses are unlikely to voluntarily share their raw data for reanalysis. In the case that authors are aware of reporting errors in their published work or suspect that the conducted data selection, data transformation, and statistical analysis are problematic in some regard, it seems even less likely that they are willing to share the underlying data. We think that it is reasonable to assume that some of the corresponding authors that received a request to share their raw data by Vanpaemel et al. (2015) reran the statistical analysis to verify them before deciding on whether or not to share their data. Authors unable to reproduce their published statistical results then potentially refrained from sharing their data. We also believe that those articles we could not reanalyze because the underlying raw data were lost, destroyed, or inaccessible at the time of the sharing request in 2013, the year following the publication, are likely to be indicative of a careless and sloppy workflow, which likely results in relatively more issues with reproducibility. How much more frequent (severe) errors are in articles where the authors are not willing (or unable) to share the raw data, however, is unknown and not directly empirically investigable.[15]

This selection effect could be quite strong, as shown by empirical evidence in the field of neuroscience. Miyakawa, then Editor-in-Chief of the journal *Molecular Brain*, asked the authors of 41 out of 180 articles that had been submitted between 2017 and 2019 in *Molecular Brain* for their raw data because the empirical findings were “too beautiful to be true” (Miyakawa, 2020). The authors of 21 of these 41 articles responded with a withdrawal of their submission without providing the raw data. The authors of the remaining 20 articles provided some raw data. However, 19 of these articles were then rejected because either the provided data was insufficient, or some of the articles’ empirical findings were found to be irreproducible. A follow-up investigation further revealed that 14 of these 41 articles were then published in other journals, 12 of which in journals that require or recommend data sharing. An email request for the raw data for these 12 articles was unanswered by 10 and declined by one. The author of the 12th article only sent some of the raw data (see Figure 1 in Miyakawa, 2020).

The second selection effect is that our study only reproduced PCs based on the most common and simple statistical models, thereby excluding more complicated and potentially more error-prone analyses. The study of Bergh et al. (2017) (see Table 1) concluded that the analyzed (complicated) SEMs had more than twice as many reproducibility issues as the analyzed (relatively simple) linear regressions. This is clear evidence for the discussed selection effect. Given its model complexity and the fact that only a small proportion of researchers are willing to share syntax or code used to fit the SEM with other researches (see Wicherts & Crompvoets, 2017), we do not suspect many strongly reproducible SEM results in the psychological literature.

Due to these two selection effects, it seems reasonable to conclude that relatively less strongly reproducible results exist in the full set of the 245 articles that were published in 2012 in one of the three analyzed journals. With 5-year impact factors in 2018 (i.e., years 2013–2017) of 3.777, 6.286, and 2.590 for *Psychology and Aging*, *Journal of Abnormal Psychology*, and *Experimental and Clinical Psychopharmacology*, respectively, the analyzed journals are representative for the majority of APA journals. Coupled with the fact that only the most basic statistical models were investigated, we conclude that it is in general very challenging to reproduce the key statistical results in articles published in APA journals in the year 2012 when the underlying raw data is made available.

In light of the above reasoning, it seems justified to assume that a randomly picked article published in an APA journal around the year 2012 has relatively more reproducibility issues compared

[15] A similar hypothesis about the relation between data sharing and consistency among test statistics, degrees of freedom, and *p* values was investigated empirically by Wicherts et al. (2011) and Nuijten et al. (2017). Their findings are compatible with either no relation or slightly more inconsistencies in articles for which data was not shared.

with a randomly picked article from our sample of reanalyzed articles. How reproducible more recent publications in APA journals are, is unclear, and can only be answered via future empirical investigations. One reason to suspect that issues with reproducibility decreased in recent years is that issues with replicability, including questionable research practices (Simmons et al., 2011) and fraud cases Stroebe et al. (2012), were put center stage since 2012. This could have potentially increased the overall care of researchers with respect to data analysis and the reporting of statistical results. Additionally, the awareness of reproducibility problems has increased and the three analyzed journals are currently encouraging authors to share their data and code of analysis together with their article (P&A and E&C) or are at least pointing out this possibility without explicitly encouraging it (JAP). In favor of generalizability of the current study to more recent publications, it turns out that in terms of explicit guidelines, no changes occurred between 2012 and 2019 since the 6th edition of the APA publication manual (APA, 2010) was in effect from 2009 until 2019. In 2020, the 6th edition of the APA publication manual was replaced with the 7th edition (APA, 2020), but this new edition did not bring meaningful changes with respect to the reporting of quantitative results.

Future Research on Reproducibility

Reproducibility crucially depends on several factors, such as the availability of software, incentives, journal policies, training, and a general awareness about the importance of and issues with reproducibility. As these factors are constantly in flux, the state of reproducibility of psychology is bound to change. Studying potential changes in the number and type of reproducibility issues in the last decade requires new empirical studies, using sufficiently large samples of more recent articles. Considering the importance of this topic, we believe that the field of psychology would be well-served by it, despite the resource-binding, time-consuming nature of such studies. Given its scope, the current study informs on typical variations within articles, within journals, and across journals and could therefore serve as a benchmark for potential future studies on the reproducibility of ANOVA and regression designs.

Future studies on reproducibility should assess both vagueness and correctness of published statistical results. One possible approach for achieving this goal is via multilab collaborations similar to the Psychological Science Accelerator (Moshontz et al., 2018) where reanalysts independently attempt to reproduce identified statistical results. Both the interrater reliability and the average time until reproduction success would then inform about the vagueness of the identified statistical results. Large-scale multilab collaborations have one distinct advantage over single lab studies like the current study, which is large manpower. This allows not only for the reproduction of a large number of articles published in a multitude of different journals and in different time periods but also for relatively quick reproducibility assessments. As a result, the effects of policy changes on the reproducibility of published results could be assessed not too long after their implementation.

Ideally, future research in psychology will explore uncharted territory when it comes to reproducibility. In particular, we would like to see reproducibility assessments of complicated and computation-heavy statistical designs such as structural equation models and multilevel VAR models as well as Monte Carlo simulation studies with their unique challenges with respect to reproducibility (see Fitzpatrick, 2019, for a discussion of these challenges).

Recommendations for Strongly Reproducible Research

Our reanalysis attempts indicate that the current workflow in psychological research frequently leads to a pain-staking reproducibility experience that relies on patient trial and error and to reporting errors. This leads us to the conclusion that the currently prevailing workflow of writing and publishing empirical articles is ill-suited to prevent irreproducible results.

Obviously, reproducibility depends on the availability of the raw data. So strictly speaking, reproducibility is much lower than 70% in this study, because for over two thirds of the articles in the three analyzed journals the key statistical results were not reproducible in principle, because the data were not made available (see Figure 1). As this study showed, sharing the study's raw data in conjunction with the article rarely suffices to make the statistical results strongly reproducible. To adequately describe the meaning of the provided data columns and how higher-level variables that are essential for the conducted analysis are then created, a codebook is necessary. For guidelines on codebooks for typical data in psychological research, the reader is referred to the codebook cookbook (see https://osf.io/72hrh/). A wonderful way to create codebooks with all the necessary metadata for common data files such as.rds,.sav,.dta,.xlsx, and.csv files can be done via the freely available *codebook* R package (Arslan, 2019). More general guidelines on what and what not to share and how to share it including detailed information on public repositories are provided by (Klein, Vianello, et al., 2018).

Likely, copy-pasting values from long statistical outputs (e.g., SPSS ANOVA outputs) containing a large number of values with test statistics, *p* values and degrees of freedom spreading over multiple pages into the article is an error-prone workflow. Journals usually take care of the final formatting of a submitted article, and we suspect that a non-negligible portion of copy-paste errors occur at this stage of the publication process both via the probabilistic reasoning that "all sorts of mistakes that can happen, happen" and by the assumption that everyone makes mistakes, including authors and copy editors of journals.

There are several easy ways to increase the probability of strongly reproducible statistical results: Avoid copy-pasting numerical values as much as possible. Round numbers only once and at the end. Avoid manual data manipulation steps, such as excluding outliers and transformations of variables in the data file. Instead, conduct data manipulations via code and store that code. To make the selection of cases transparent in your article, use flowcharts as they are custom in the biomedical literature (see, e.g., the CONSORT guidelines http://www.consort-statement.org/). Doing so maximizes transparency and facilitates the ease of performing alternative data transformations without the need to store multiple data files that differ only slightly. Ask a colleague to redo your analyses based on your description, which provides useful information both about the correctness of your numerical result and about the clarity of your reporting (and return the favor later). Such independent reanalyses are part of the "co-pilot model of statistical analysis" (see Veldkamp et al., 2014; Wicherts, 2011). Further, share the analysis code, as it is a great way to remove interpretation issues regarding the specifics of the fitted model. You can easily create the analysis code in any standard statistical software even when using a graphical user interface when analyzing the data (e.g., by clicking "Paste" as opposed to "OK" in SPSS). Beware that it can be incredibly hard to comprehend many lines of code. By adding comments to the code of analysis, you aid other researchers that want to reproduce your results considerably. Naturally, the output generated by computer code depends on all the software that is being used to run the code. For a discussion of the intricacies of generating reproducible statistical results when using continuously changing statistical software, including code containers like Code Ocean, see Epskamp (2019) and the references therein.

Probably the most potent way to increase reproducibility is by writing a so-called dynamic report (document/notebook). A dynamic report is the combination of the article's narrative (i.e., all the text) and computer code for all statistical analyses that are directly linked to the raw data. The converted (rendered) report then comprises the narrative, as well as all code and generated output (e.g., summary statistics, results of statistical analyses, figures) that one wishes to show to the reader. This converted report then constitutes an article ready for publication, and because of the connection of raw data, statistical analysis, and output, the likelihood of incorrect statistical results as a result of copy-and-paste, repeated rounding, or intermediary rounding, is drastically reduced.

In the case that a dynamic report is published in conjunction with the article, the vagueness of the statistical results in the article is drastically reduced. Inconsistencies between the selections (e.g., the applied exclusion criteria), transformations (e.g., the computation of some index), and methods (e.g., the used estimator) described in the article's narrative and the actually executed calculations can be identified by the reader rather handily by looking at the code and its connection with the produced statistical output. The better structured and the clearer the code, the easier is such a consistency check. To achieve maximal transparency and to enable the reader to conduct a reproducibility check for every statistical result in the rendered report, the used raw data sets have to be made available in addition to the dynamic report, ideally sufficiently supplemented with metadata that makes clear what each column and all their entries represent. Ideally, a dynamic report includes information about the necessary version(s) of the used statistical software, packages, and functions. In R (R Core Team, 2018) this can be done via the command sessionInfo().

The statistical programming language R (R Core Team, 2018) currently offers two ways to create dynamic reports, via LATEX code (Sweave) and via markdown code (Rmarkdown) (see RStudio Team, 2015). Writing a dynamic document with the LATEX language offers an extensive range of options, but it takes some time to become efficient. Getting acquainted with the Markdown language, on the other hand, is quite easy and for most purposes, its range of possibilities should be sufficient.[16] Another option for creating dynamic reports are Jupyter notebooks, which can be used in conjunction with R (R Core Team, 2018) or Python (Python Core Team, 2015).

For researchers who are not keen on writing code but prefer to work with a graphical user interface instead, we recommend exhausting all transparency options offered by the used software. For example, users of the freely available statistical software JASP (JASP Team, 2020) could share the .jasp file used to generate the statistical analyses together with the article, as this would reveal which variables and cases were selected, the fitted statistical model, as well as the chosen estimators. Sufficient annotation then allows other researchers to link results in the article with those in the .jasp file. To disseminate the used data files and the .jasp file, researchers can, for instance, use the Open Science Framework (https://osf.io/). In doing so, it is advisable to make use of the OSF's version control feature to track changes made to the used raw data or the performed analyses.

Conclusion and a Look Forward

Although discussions about reproducibility issues in psychology are certainly not new (see e.g., Wolins (1962)), systematic empirical investigations are rare. Our study indicates, together with the few other existing empirical investigations of reproducibility (see Table 1), that it is imprudent just to assume that reporting errors are rare and noninfluential in psychological research, let alone that reported numbers are strongly reproducible with just the article and the raw data. However, empirical evidence does not suggest that reproducibility issues are particularly bad in the field of psychology. Potentially, all empirical research in the social sciences is subverted with irreproducible statistical results (see Stodden, 2015, for a selection of infamous examples of irreproducibility).

The time-consuming, painstaking nature of validating published statistical results without a code of analysis also means that we will never be able to validate large numbers of articles published in the past—even in the case that the relevant raw data is made available by the authors. Moving forward as a science, strongly reproducible results simply have to become the norm in psychology. Delightfully, some journals and initiatives have already started to emphasize this issue: The Center for Open Science (Nosek et al., 2012) and curatescience.org (LeBel et al., 2018), for instance, are currently working on reproducibility badges, signaling the crowdsourced verification of all relevant results by independent researchers, and the journal *Metapsychology*, which conducts reproducibility check on all statistical analyses in the submitted article, already has one. In the case of *Metapsychology*, the reproducibility checks are conducted by a team consisting of the statistical editor and the editorial assistant. Such an in-house review constitutes an excellent service to the scientific community. Due to the resource-consuming nature of such an approach, it does not seem likely that large journals that publish multiple issues of empirical articles per year will follow suit. An approach that does seem feasible for all scholarly journals is only to reproduce some empirical findings per article or only to reproduce a random sample of submitted articles (see Stodden et al., 2016). An alternative approach would be to burden the reviewers of an article with that task. This could be done by splitting duties, where a methodologically skilled reviewer focuses on the correctness and transparency of the conducting empirical analyses. In contrast, another reviewer focuses on the general quality of the article. It seems paramount to add that any rigorous reproducibility check will be time-consuming, even if the code of analysis was shared with the raw data. The reason is that it does not suffice only to check whether running the code produces the results reported in the data. Instead, it also has to be checked whether the code adequately does what is described in the Method section of the article.

[16] An R package is currently being developed that will facilitate the creation of documents that comply with the APA guidelines using Rmarkdown (Aust & Barth, 2018).

Making research transparent and reproducible by sharing desensitized raw data and code of analysis, ideally in the form of dynamic reports, is a powerful way to decrease the chances of reporting errors of honest researchers and to allow the efficient reuse of gathered data. One very ambitious project that will hopefully find imitators in psychology is the Reproducible Document Stack (RDS) project of the online journal *eLife*. This project aims at publishing articles in the form of reproducible documents in a way that establishes strong reproducibility while simultaneously allowing the reader to interact with the code of analysis and the data to generate alternative figures, tables, and statistical analyses (see, e.g., https://repro.elifesciences.org/example.html) in a user-friendly way.

Because sharing of well-documented data together with well-commented code of analysis and all necessary metadata constitutes a great service to the scientific community, we agree with Stodden et al. (2016) who advocate that digital scholarly objects such as data sets stored on a third-party repository should be cited in the reference section to credit its distributors adequately. Any future research that makes use of such a digital scholarly object (e.g., by data mining an existing dataset or by using an R (R Core Team, 2018) package that was created in light of a scientific publication) should then equally cite that object in the reference section. Such a system would incentivize the sharing of data sets, code of analysis, algorithms, and software and, thereby, assist in the quest for strongly reproducible research.

As is demonstrated in a witty and delightfully funny way in McCullough et al. (2006), sharing of data together with some code of analysis that was used does not necessarily allow the reproduction of the empirical findings. What is further needed is all necessary metadata, including information about the raw data, on how to run the code and the software requirements. To allow for a complete, computer-assisted reproduction of empirical findings Wilkinson et al. (2016) developed four elaborate principles for digital scholarly articles. These principles are findability, accessibility, interoperability, and reusability. Scholarly articles that follow all of these principles can then be conveniently labeled as FAIR.

Another advantage of a scientific shift toward transparency and strongly reproducible results is that it can potentially make the lives of science fraudsters, who deliberately omit, delete, change, or make up raw data or statistical analyses with bad intentions, considerably harder.

We hope that an increasing number of researchers in psychology will take full advantage of all the transparency options that technological progress has brought us. We believe that providing the necessary tools to make statistical analyses maximally transparent and minimally error-prone, while simultaneously stressing the importance of openness in scientific research to the new generation of scientists, will change the field of psychology for the better.

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*, *2*(2), 169–187.

American Psychological Association. (2010). *Publication manual of the American Psychological Association*.

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.).

Aust, F., & Barth, M. (2018). *papaja: Prepare reproducible APA journal articles with R Markdown* (Version 0.1.0.9842) [Computer software]. Retrieved from https://github.com/crsh/papaja

Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678.

Bakker, M., & Wicherts, J. M. (2014). Outlier removal and the relation with reporting errors and quality of psychological research. *PLoS ONE*, *9*(7), e103360.

Bergh, D. D., Sharp, B. M., Aguinis, H., & Li, M. (2017). Is there a credibility crisis in strategic management research? evidence on the reproducibility of study findings. *Strategic Organization*, *15*(3), 423–436.

Berle, D., & Starcevic, V. (2007). Inconsistencies between reported test statistics and p-values in two psychiatry journals. *International Journal of Methods in Psychiatric Research*, *16*(4), 202–207.

Brown, N. J., & Heathers, J. A. (2017). The grim test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, *8*(4), 363–369.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644.

Caperos, J. M., & Pardo Merino, A. (2013). Consistency errors in p-values reported in Spanish psychology journals. *Psicothema*, *25*(3), 408–414.

Chang, A., & Li, P. (2015). *Is economics research replicable? sixty published articles from thirteen journals say 'usually not'*. Board of Governors of the Federal Reserve System Finance and Economics.

Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in empirical economics: The journal of money, credit and banking project. *The American Economic Review*, *76*, 587–603.

Epskamp, S. (2019). Reproducibility and replicability in a fast-paced methodological world. *Advances in Methods and Practices in Psychological Science*, *2*(2), 145–155.

Epskamp, S., & Nuijten, M. (2018). *statcheck: Extract statistics from articles and recompute p-values* (R package version 1.3.1). Retrieved from https://cran.r-project.org/web/packages/statcheck/index.html

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

Fitzpatrick, B. G. (2019). Issues in reproducible simulation research. *Bulletin of Mathematical Biology*, *81*(1), 1–6.

Freese, J. (2007). Replication standards for quantitative social science: Why not sociology? *Sociological Methods & Research*, *36*(2), 153–172.

García-Berthou, E., & Alcaraz, C. (2004). Incongruence between test statistics and p values in medical articles. *BMC Medical Research Methodology*, *4*(1), 13.

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time* [Unpublished manuscript]. Department of Statistics, Columbia University, New York, NY.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12-341ps12.

Hardwicke, T. E., Mathur, M. B., Macdonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Mohr, A. H., Clayton, E., Yoon, E. J., Tessler, M. H., 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.

Ioannidis, J. P., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T., Nitzberg, M., Page, G. P., Petretto, E., & van Noort, V. (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics*, *41*(2), 149–155.

JASP Team. (2020). *JASP* (Version 0.13.1) [Computer software]. Retrieved from https://jasp-stats.org/

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Hofelich Mohr, A., Ijzerman, H., Nilsonne, G., Vanpaemel, W., & Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, *4*(1), 1–15.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, S., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., . . . 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.

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*, *1*(3), 389–402.

Maassen, E., van Assen, M. A., Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PloS ONE*, *15*(5), e0233107.

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed). Routledge.

McCullough, B. D., McGeary, K. A., & Harrison, T. D. (2006). Lessons from the JMCB archive. *Journal of Money, Credit and Banking*, *38*(4), 1093–1107.

Miyakawa, T. (2020). *No raw data, no science: Another possible source of the reproducibility crisis*. BioMed Central.

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of Bayes Factors for Common Designs* (Version 0.9.12-2) [Computer software]. Retrieved from https://cran.r-project.org/src/contrib/Archive/BayesFactor/BayesFactor_0.9.12-2.tar.gz

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., . . . Chartier, C. R. (2018). The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515.

Naudet, F., Sakarovitch, C., Janiaud, P., Cristea, I., Fanelli, D., Moher, D., & Ioannidis, J. P. A. (2018). Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: Survey of studies published in The *BMJ* and *PLoS* Medicine. *British Medical Journal*, *360*, k400–k411.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: Ii. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

Nuijten, M. B., Borghuis, J., Veldkamp, C. L., Dominguez-Alvarez, L., van Assen, M. A., & Wicherts, J. M. (2017). Journal data sharing policies and statistical reporting inconsistencies in psychology. *Collabra: Psychology*, *3*(1), 1–22.

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior research methods*, *48*(4), 1205–1226.

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

Petrocelli, J. V., Clarkson, J. J., Whitmire, M. B., & Moon, P. E. (2013). When ab≠ c–c’: Published errors in the reports of single-mediator models. *Behavior research methods*, *45*(2), 595–601.

Pollock, J. J., & Zamora, A. (1983). Collection and characterization of spelling errors in scientific and scholarly text. *Journal of the American Society for Information Science*, *34*(1), 51–58.

Python Core Team. (2015). *Python: A dynamic, open source programming language* [Computer software]. Retrieved from https://www.python.org/

R Core Team. (2018). *R: A language and environment for statistical computing* [Computer software]. Retrieved from https://www.R-project.org/

Rosnow, R. L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *57*(3), 221.

Rossi, J. S. (1987). How often are our statistics wrong? A statistics class exercise. *Teaching of Psychology*, *14*(2), 98–101.

RStudio Team. (2015). *Rstudio: Integrated development environment for r* [Computer software]. Retrieved from http://www.rstudio.com/

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149.

Stodden, V. (2015). Reproducing statistical results. *Annual Review of Statistics and Its Application*, *2*, 1–19.

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., Heroux, M. A., Ioannidis, J. P. A., & Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science*, *354*(6317), 1240–1241.

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, *7*(6), 670–688.

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? the availability of psychological research data after the storm. *Collabra: Psychology*, *1*(1), Article 3.

Veldkamp, C. L., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PloS ONE*, *9*(12), e114876.

Wicherts, J. M. (2011). Psychology must learn a lesson from fraud case. *Nature*, *480*(7375), 7–7.

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.

Wicherts, J. M., & Crompvoets, E. A. (2017). The poor availability of syntaxes of structural equation modeling. *Accountability in research*, *24*(8), 458–468.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). The FAIR Guidiing Principles for scientific data management and stewardship. *Scientific Data*, *3*(160018), 1–9.

Wolins, L. (1962). Responsibility for raw data. *American Psychologist*, *17*(9), 657–658.

Xie, Y. (2015). *Dynamic documents with r and knitr* (Vol. 29). Chapman & Hall/CRC.

Received June 4, 2019
Revision received August 31, 2020
Accepted September 4, 2020 ■