Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy

Anna Laurinavichyute
Department of Linguistics, University of Potsdam, Potsdam, Germany

Himanshu Yadav
Department of Linguistics, University of Potsdam, Potsdam, Germany

Shravan Vasishth
Department of Linguistics, University of Potsdam, Potsdam, Germany

April 11, 2022

Author Note

Please send correspondence to anna.laurinavichyute@uni-potsdam.de.

Abstract

In 2019 the Journal of Memory and Language instituted an open data and code policy; this policy requires that, as a rule, code and data be released at the latest upon publication. How effective is this policy? We compared 59 papers published before, and 59 papers published after, the policy took effect. After the policy was in place, the rate of data sharing increased by more than 50%. We further looked at whether papers published under the open data policy were reproducible, in the sense that the published results should be possible to regenerate given the data, and given the code, when code was provided. For 8 out of the 59 papers, data sets were inaccessible. The reproducibility rate ranged from 34% to 56%, depending on the reproducibility criteria. The strongest predictor of whether an attempt to reproduce would be successful is the presence of the analysis code: it increases the probability of reproducing reported results by almost 40%. We propose two simple steps that can increase the reproducibility of published papers: share the analysis code, and attempt to reproduce one's own analysis using only the shared materials.

Keywords: open data; reproducible statistical analyses; reproducibility; open science; meta-research; journal policy

Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy

Introduction

Being able to build on existing knowledge is key to scientific progress. A prerequisite for such incremental knowledge gain is that existing research be reliable, in the sense that we can be relatively confident that the claimed findings in a paper reflect something that is true about the phenomenon being investigated. The replicability of published claims is a key component of reliability (Munafò et al., 2017; National Academies of Sciences & Medicine, 2019). Here, replicability can be broadly understood to mean that one can obtain the same or similar conclusions as in a published result when one repeats the same experiment with new participants.[1]

In recent years, the replicability of apparently well-established results in psychology has been called into question (e.g., Anderson et al., 2015). Since then, within psychology and psycholinguistics several failed replication attempts of well-known claims have been reported (e.g., Hagger et al., 2016; Jäger et al., 2020; R. A. Klein et al., 2014; R. A. Klein et al., 2018; Mertzen et al., 2020; Nieuwland et al., 2020; Nieuwland et al., 2018; Stack et al., 2018; Vasishth et al., 2018). Such failures to replicate raise important questions about the extent of the non-replicability problem in psycholinguistics and related areas. Partly in response to this concern, a special issue of the *Journal of Memory and Language* focused on the replication of influential findings in memory and language research. This move from the journal is a clear signal to the field that replication has an important role to play in scientific progress.

Replicability is widely regarded as an important aspect of assessing the reliability of a particular finding. However, if one fails to replicate a study, it is often difficult to work out

[1] By "same or similar" we are not referring to statistical significance/non-significance but rather to the broad consistency of estimates across replications (for detailed discussion about consistency of observed patterns in the psycholinguistic context, see Vasishth & Gelman, 2021).

exactly why the failure to replicate happened. This is because the failure may be driven by many factors that are external to the research question under investigation: differences in the population and/or language studied, the natural variability in the dependent variable, lab settings, equipment, and protocols can come together to lead to very different outcomes. Indeed, although large effects are generally easy to replicate (an example is certain garden path constructions, Paape & Vasishth, 2022), when it comes to studying subtle and highly variable aspects of human behavior, replicability may well be an unattainable goal. This inherent variability of effects is why statisticians like Andrew Gelman often emphasize the need to embrace variation and accept uncertainty (https://youtu.be/E8uzPjg1mR8).

Thus, the replicability of any particular finding could be beyond the control of the researcher because of the type of phenomenon being studied or the inherent variability in the behavioral response. However, there is another important aspect of reliability that is directly under the researcher's control. This fundamental aspect is the reproducibility of published claims. Reproducibility refers to the ability to regenerate the key summary statistics (e.g., means, standard errors, t-, F-, or p-values) reported in a paper using the data (and code) provided with the paper (LeBel et al., 2018; Nuijten et al., 2018; Stodden et al., 2018). Given some fixed data, if the key summary statistics that the statistical inference is based on are not reproducible, it is unclear what a replication attempt should even aim to replicate.

On the face of it, reproducibility might seem to be an easily attainable goal: is one not bound to obtain the same results on the original data set if one simply re-runs the data analysis code again? In recent years, researchers in psychology and other areas have come to learn that success in reproducing the original results is not a given, mainly due to the absence of data and code. Researchers in different fields have been investigating the reproducibility of published claims. Some examples are psychology (Nuijten et al., 2016), political science (Stockemer et al., 2018), economics (Chang & Li, 2015), biomedical science (Naudet et al., 2018), machine learning (Raff, 2019), and hydrology (Stagge et al., 2019). Across all these studies, the reproducibility rate has been reported to be around 30%,

ranging from 17% in political science (Eubank, 2016) to 37% in economics (Chang & Li, 2015), with some exceptions (58% for registered reports in psychology, Obels et al., 2020; 70% for articles whose authors shared data upon request, Artner et al., 2021; 55% for meta-analyses in psychology, Maassen et al., 2020; 82% in biomedicine, Naudet et al., 2018). In recognition of this irreproducibility crisis, the journal *Cortex* has introduced a new article type, the verification report. The sole purpose of this article type is to repeat the original analyses or report new analyses of the original data set (Chambers, 2020).

One of the most basic barriers to reproducibility is data unavailability: if the original data set is not openly available (which is often the case, see Vanpaemel et al., 2015; Wicherts et al., 2006), independent analysis cannot be performed. In response to this challenge, many journals in psychology and linguistics have adopted a mandatory data sharing policy. Examples are *Cognition*, *Cortex*, *Collabra:Psychology*, *Open Mind*, *Glossa Psycholinguistics*, *Journal of Cognition*, and *Computational Brain and Behavior*. More than a thousand journals from a wide range of fields are signatories to the Transparency and Openness Promotion (TOP) guidelines from the Center for Open Science (https://www.cos.io/initiatives/top-guidelines).

The introduction of such guidelines and policies has been effective in increasing the proportion of open data sets accompanying articles (Hardwicke et al., 2018; Nuijten et al., 2017). However, although ensuring data availability is a necessary condition for reproducibility, it may not be sufficient: Towse et al. (2021) analyzed the quality of open data sets across psychological journals. They report that 51% of the surveyed data sets were not sufficiently documented or did not contain all the data needed to reproduce the reported analyses, and 68% were archived in a way that limits reusability. Examples of non-reusability were data that were not machine-readable, aggregated, or missing (for a similar point, see also Hardwicke et al., 2021).

The *Journal of Memory and Language* was among the first linguistics-oriented journals to adopt, in 2019, the mandatory data sharing policy, thus making the first step on the path

to ensuring reproducibility of published findings (Gerrig & Rastle, 2019). To assess the impact of this policy, in May 2020, the editor-in-chief of the *Journal of Memory and Language* (Professor Kathleen Rastle) commissioned an independent evaluation of whether the rate of data sharing has increased and whether the shared data is sufficient to reproduce the published results. The aim of the present work is to evaluate the data sharing policy and to investigate the reproducibility of papers published in the *Journal of Memory and Language* after this policy was instituted.

Evaluation of the reproducibility of JML articles (2019–2021)

Open data policy. When submitting a manuscript to the Journal of Memory and Language, on the first page of the submission form, the authors are confronted with the open data policy:

> "We require articles published in the Journal to make publicly available any stimuli, data, analysis code, and computational models associated with the research. Because these materials are an important part of the research that the Journal is reviewing, we require authors to provide a (private) link to them at the point of submission. Please include this link on the title page of your manuscript. Manuscripts that do not include access to these materials will usually be returned to authors.
>
> Further information about our data sharing policy is available here."

Material. The editor-in-chief gave us the titles of papers reporting quantitative experiments that were published before (N=59) and after (N=59) data sharing was made mandatory in JML in 2019.

The list of papers published before the open data policy took effect: Akan et al. (2018), Arnold et al. (2018), Chan et al. (2018), Chubala et al. (2018), Cunnings and Sturt (2018), de Bruin et al. (2018), Deliens et al. (2018), Do and Kaiser (2019), Drummer and Felser

(2018), Fisher and Radvansky (2018), Frazier and Clifton Jr (2018), Fritz et al. (2019), Fukumura (2018), Galati et al. (2019), Gathercole et al. (2019), Healey (2018), Hopper and Huber (2018), Hsiao and Nation (2018), Isarida et al. (2018), James et al. (2018), Jones and Farrell (2018), Jou et al. (2018), Karimi et al. (2018), Keung and Staub (2018), Kowialiewski and Majerus (2018), Malejka and Bröder (2019), McKoon and Ratcliff (2018), Miller et al. (2019), Miyoshi et al. (2018), Mohanty and Naveh-Benjamin (2018), Nicenboim and Vasishth (2018), Nooteboom and Quené (2019), Osth et al. (2018), Paap et al. (2019), Sahakyan and Malmberg (2018), Scott and Sera (2018), Seabrooke et al. (2019), Seedorff et al. (2018), Singh et al. (2018), Slioussar (2018), Stefanidi et al. (2018), Susser and Mulligan (2019), Thalmann et al. (2019), Uner and Roediger III (2018), Van Bergen and Bosker (2018), van Heugten et al. (2018), van Tiel et al. (2019), Vasishth et al. (2018), Vaughn and Kendall (2018), Veldre and Andrews (2018a, 2018b), Wang et al. (2018), Wedel et al. (2018), Wen and van Heuven (2018), B. M. Wilson et al. (2019), Yim et al. (2018), Zawadzka et al. (2018), Zhang and Samuel (2018).

The list of papers published after the open data policy took effect: Ahn and Brown-Schmidt (2020), Ambrus et al. (2020), Avetisyan et al. (2020), Bangerter et al. (2020), Boyce et al. (2020), Brainerd, Bialer, et al. (2020), Brainerd et al. (2019), Brainerd, Nakamura, and Murtaza (2020), Brandt et al. (2020), Brewer et al. (2021), Bristol and Rossano (2020), Brothers and Kuperberg (2021), Brysbaert (2019), Bürki et al. (2020), Chan et al. (2020), Chetail (2020), Collins et al. (2020), Corps and Rabagliati (2020), Dıéez-Álamo et al. (2020), Falandays et al. (2020), Fellman et al. (2020), Floccia et al. (2020), Fox et al. (2020), Fujita and Cunnings (2020), Gagné et al. (2020), Garnham et al. (2020), Günther, Nguyen, et al. (2020), Günther, Petilli, et al. (2020), Hesse and Benz (2020), Hollis (2020), Humphreys et al. (2020), Hwang and Shin (2019), Isarida et al. (2020), Jäger et al. (2020), Johns et al. (2020), Kaula and Henson (2020), Lange et al. (2019), Lauro et al. (2020), Lelonkiewicz et al. (2020), Li et al. (2020), Liang et al. (2021), McKinley and Benjamin (2020), Meteyard and Davies (2020), Nooteboom and Quené (2020), Osth et al. (2020),

Reifegerste et al. (2020), Saito et al. (2020), Samuel (2020), Schubert et al. (2020), Siegelman et al. (2020), Siew et al. (2021), Skrzypulec and Chuderski (2020), Snefjella et al. (2020), Snell and Theeuwes (2020), Tirso and Geraci (2020), Troyer and Kutas (2020), Tsuboi and Francis (2020), Villani et al. (2021), Yang et al. (2020).

Data Availability

The code and anonymized data for regenerating this paper are available from https://osf.io/3bzu8/.

The detailed anonymized attempts to reproduce the 59 papers published under open data policy are available from https://osf.io/3x2y6/. The table establishing the correspondence between anonymized IDs and paper titles can be found at the root folder of the same project.

Methods

We downloaded any available material[2] published with every paper. Data were labeled as accessible when at least some subset (but not necessarily all) of the data described in the paper had been made available. We additionally annotated whether the code for performing the analysis was present, the analysis was preregistered, and whether data was accompanied by a readme file or the like explaining what variable names and their values in the data file stand for. These factors could potentially affect reproducibility: the analysis code documents all the analysis steps; the readme file identifies the explanatory and dependent variables; and a preregistration may be associated with a more reproducible analysis code (for example, Obels et al. (2020) reported a relatively high reproducibility rate, 58%, for registered reports in psychology).

For every paper published under the open data policy, we attempted to reproduce the analysis using the description in the paper itself, and the analysis code, if this was provided.

[2] For one paper in the post-policy list and nine papers in the pre-policy list, materials were published on the journal website; they were forwarded to us by the editor-in-chief.

When a reported value could not be reproduced, alternative computations compatible with the analysis description were attempted. For example, when the reported condition mean could not be reproduced with data aggregated within participants, we computed the mean without the aggregation step. When the difference between the reproduced and the reported values was greater than 10% for at least 20 reported values, the attempt to reproduce was terminated. If data for some of the reported experiments or analyses was missing, we attempted to reproduce the remaining results. In contrast to some reproducibility assessments (Hardwicke et al., 2021; Hardwicke et al., 2018), we did not contact the authors and ask for clarifications if the results could not be reproduced. This decision follows from the goal of computational reproducibility: to obtain the same results using available data and procedures. If not all analysis steps were described in sufficient detail, then the study is not reproducible for the reader, even if these analysis steps are fully documented for private use. The outcomes of the reproducibility attempts were evaluated under several criteria, ranging from a strict to several increasingly relaxed ones.

Strict criterion: Papers were labeled as reproducible if all the reported analyses could be reproduced exactly, including reported means (except for the rounding errors and the edges of Bayesian credible intervals, where minor fluctuations are expected). If we could not reproduce the exact numbers, even when all the results reported as significant remained significant, the paper was considered to be not reproducible. The reason we focus on the reproducibility of the summary statistics and not whether an effect was significant or not is that statistical significance per se is not a very informative result, unless the power properties of the experimental design are also known (Gelman & Carlin, 2014; Jäger et al., 2017; Jäger et al., 2020; Vasishth & Gelman, 2021; Vasishth et al., 2018).

The exact effect estimates from an analysis could be not reproducible for a number of reasons, including updates to the software used to run the analysis, using a updated version of the optimizer for fitting a linear mixed model, etc. Note, however, that summary statistics such as means and standard deviations, do not depend on the software. Obviously, the fact

that a study was not reproducible under the strict criterion does not necessarily say anything about the quality of scientific evidence presented in the study, but rather about the quality of data and analysis presentation.

While being straightforward and easy to evaluate, the strict reproducibility criterion may be seen as too mechanical: it does not necessarily correspond to researchers' intuition about reliable results. Most researchers would probably still consider results to be reliable even if several reproduced values slightly deviated from the reported ones. For this reason, we introduced a set of more relaxed criteria that may better correspond to the intuitive notion of reproducibility.

Relaxed criteria: If there are less than K cases where the reproduced value differs from the reported value by a margin of more than 10%, the study is considered reproducible under a relaxed criterion. Discrepancies smaller than 10% of the reported value are disregarded. The "10% criterion" subsumes all possible discrepancies: neither p-values nor missing data receive special treatment. If data for some experiment is missing, all the values that are reported for this experiment are labeled as irreproducible by a margin of more than 10%.

This criterion is evaluated at several Ks: 1, 5, 10, and 20. For example, we find that we cannot reproduce four values by a margin of (more than) 10%: in four cases, the reproduced value differs from the reported value by a margin of over 10%. In this situation, we say that the study is reproducible under K=5 criterion, but not under K=1 criterion.

The above range of reproducibility criteria (K: 1 to 20) was chosen based on the following rationale. The lower threshold, a single major discrepancy blocking reproducibility, corresponds to the criterion used in the reproducibility assessments by Hardwicke et al. (2021), Hardwicke et al. (2018). The upper threshold roughly corresponds to a small experiment being irreproducible, with major discrepancies in both condition means and inferential statistics. We suggest this as the upper threshold for discrepancies in papers that could still be considered reproducible in a broad sense.

We do not evaluate whether the main claims of the study hold: identifying which

results correspond to the main claims is a nuanced decision that is not always within our expertise. For each study, the reader can make their own informed decision based on the published attempt to reproduce the results.

Descriptive statistics

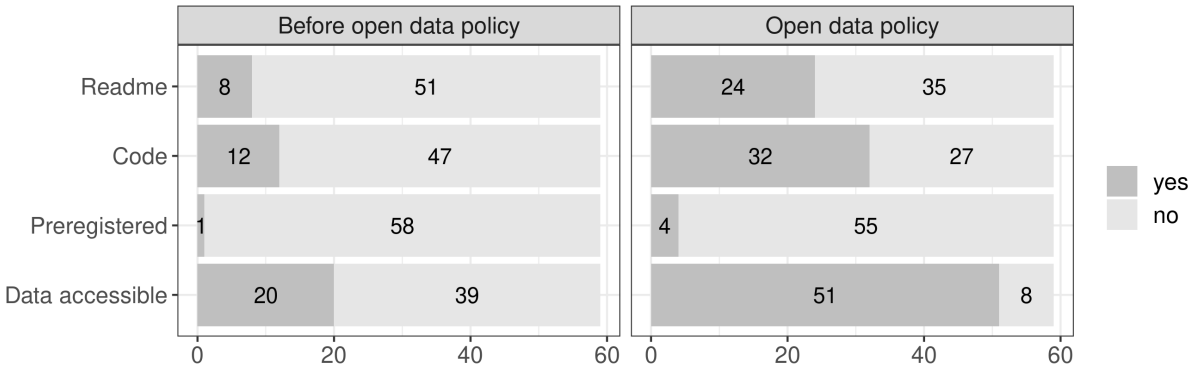

Figure 1. A summary showing the number of papers which had a readme file or the like that provided some documentation, for which the code was available, which were preregistered, and for which the data were accessible.

The key properties of the data and code release that may affect reproducibility are summarized in Figure 1. Open data policy has increased the rate of data sharing by more than 50%: 20 (34%) papers in the pre-policy group vs. 51 (86%) papers in the post-policy group had downloadable data sets. While data sharing in the pre-policy group was voluntary, in the post-policy group, the authors were required to make the data openly available. Still, the data were inaccessible for 8 papers. The reasons for data unavailability were the following:

- Four papers had no link to data/computational model introduced in the paper.
- For three papers, the data cannot be retrieved from the link provided. The reasons were diverse: the page does not exist (N=1), the link leads to a university repository that requires a log-in using the university account (N=1), supplementary materials contain a description of additional analyses instead of data (N=1).

- One paper stated that "data will be available upon request"; our request was met with a question about the planned use for the requested data. We did not reveal that we were carrying out a reproducibility check, and marked this data set as inaccessible, since the journal policy states that data should be made available unconditionally.

In the pre-policy group, of 20 papers with accessible data, 12 (60%) also shared analysis code. Most papers (18 of 20, 90%) shared all the data necessary to reproduce the reported analyses, that is, all independent and dependent variables reported in the manuscript were present in the data file.

In the post-policy group, of 51 papers with accessible data, 32 (63%) also shared analysis code. Only 38 (75%) shared all the data necessary to reproduce the reported analyses.

Results

Strict criterion. Of 19 papers that shared data but not analysis code, only 1 (5%) could be fully reproduced. Of 32 papers that shared both data and analysis code, 19 papers (59%) could be fully reproduced. Overall, 20 papers could be fully reproduced, which constitute 39% of all papers that had accessible data, and 34% of the 59 surveyed papers. Figure 2A summarizes the breakdown of reproducible papers depending on the presence of analysis code.

Relaxed criteria. The outcomes of the reproducibility assessment under increasingly relaxed criteria are summarized in Table 1. For the remainder of this paper, we will focus on the most lenient criterion that tolerates up to 20 major discrepancies between the reported and the reproduced values. Under this criterion, the number of reproducible papers that shared data but not analysis code increased to 7 of 19 (37%). Of 32 papers that shared both data and analysis code, 26 papers (81%) could be reproduced. Overall, 33 papers could be fully reproduced, which constitute 65% of all papers that had accessible data, and 56% of the 59 surveyed papers. Figure 2B summarizes the breakdown of reproducible papers

depending on the presence of analysis code.

Table 1
The number of papers classified as reproducible according to different reproducibility criteria.

	Reproducible papers	
	Without code (of 19)	With code (of 32)
Strict reproducibility criterion	1	19
Only minor discrepancies	1	20
Up to 5 major discrepancies	3	22
Up to 10 major discrepancies	4	23
Up to 20 major discrepancies	7	26

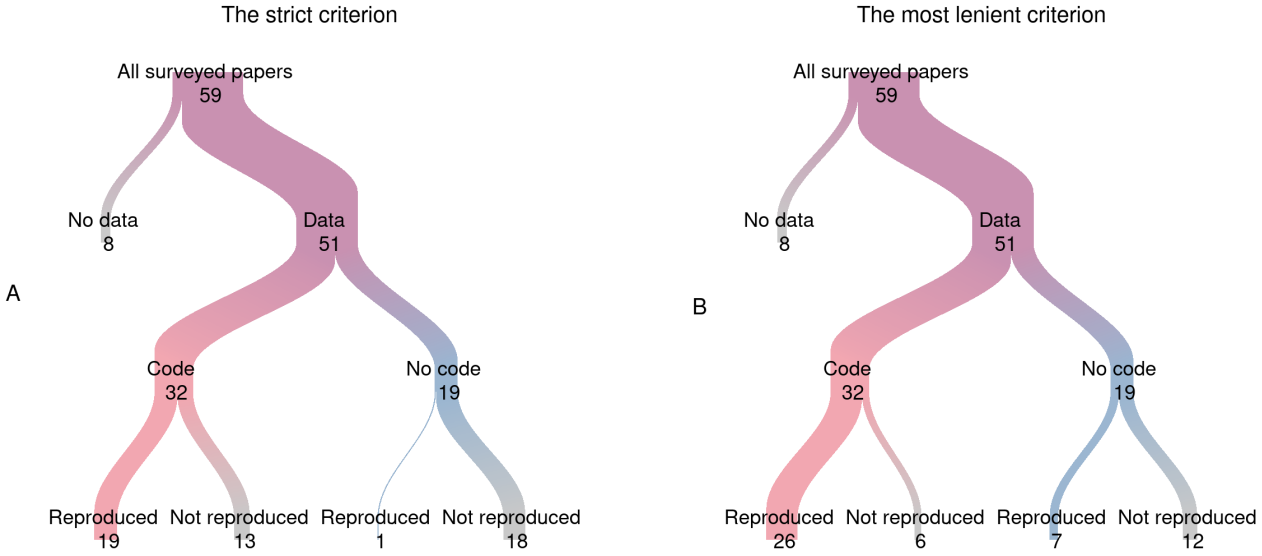

Figure 2. Summary of the reproducibility rates according to the strict (panel A) and the most lenient (panel B) criteria of reproducibility.

Modeling. We investigated whether the probability of reproducing an analysis (according to both the strict and the most lenient criterion) is affected by the availability of code, a readme file or the like, and whether the study was preregistered. This analysis was carried out via a logistic multiple regression using the `brms` package (Bürkner, 2017). The absence of a particular feature was coded as 0, and presence of that feature as 1, so that the intercept of the model corresponds to the estimated probability of reproducing the analysis

for which only data, but not code, readme, or preregistration are available. A $\mathcal{N}(0, 1.2)$ regularizing prior (Nicenboim et al., 2021) was defined for the intercept; this allows the prior probability of reproducibility given no documentation, no preregistration, and no code to lie anywhere between 4% and 96% with 95% probability (mean: 50%). The prior for the slopes was $\mathcal{N}(0, 1.5)$; this implies that the prior probability of how reproducibility could change when analysis code is present lies between -55% and 54% with 95% probability (mean: 0%).

Strict criterion. The outcomes of the analysis are shown in Table 2: the estimated probability of reproducing analysis of a paper for which only the data set is shared (intercept) is 7% with a Bayesian 95% credible interval [1, 22]%. While both readme file and preregistration seem to increase the probability of reproducing an analysis, these estimates are very uncertain and also allow for 0 and negative effects, which means that the available data do not allow us to quantify whether and how these factors affect reproducibility. In contrast, the availability of code is associated with a big increase in probability of reproducing an analysis: the estimated increase due to the availability of code is 38% with 95% CrI [19, 60]%.

Table 2
Estimates of the log odds of reproducing an analysis according to the strict reproducibility criterion: estimating the contribution of analysis code, readme, and preregistration.

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	−2.61	0.76	−4.24	−1.29
Readme	0.72	0.62	−0.47	1.93
Preregistered	0.98	1.02	−0.96	3.01
Code	2.49	0.76	1.11	4.08

The most lenient criterion. The outcomes of the analysis are shown in Table 3: compared to 7% under the strict criterion, the estimated probability of reproducing analysis for which only the data set is shared is increased to 39% with a Bayesian 95% credible interval [19, 62]%. Interestingly, the availability of code is associated with the same increase in probability of reproducing an analysis as under the strict criterion: the estimated increase is 39% with 95% CrI [14, 60]%.

Table 3
Estimates of the log odds of reproducing an analysis according to the most lenient reproducibility criterion: estimating the contribution of analysis code, readme, and preregistration.

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	−0.46	0.51	−1.47	0.51
Readme	0.05	0.61	−1.18	1.23
Preregistered	0.28	0.99	−1.61	2.31
Code	1.76	0.60	0.63	2.96

Discussion

Although the reproducibility rate according to the strict criterion (34%) may seem discouraging, our estimate is not very different from those reported in other reproducibility attempts that used similarly strict reproducibility criteria (Table 4). No reproducibility attempts used exactly the same criteria, so the outcomes cannot be directly compared. Nevertheless, it is useful to see what the numbers are like across such attempts, keeping in mind that there are important differences. Stodden et al. (2018), who analyzed the reproducibility of papers published in *Science* after the data sharing policy took effect, report an estimated reproducibility of 26%. Hardwicke et al. (2018), who assessed the reproducibility of papers published in *Cognition*, estimated a reproducibility rate of 31% without author assistance, and an additional increase of 31% when the authors of the original manuscript helped to reproduce the outcomes. Similarly, Hardwicke et al. (2021) report a reproducibility rate of 36% without author assistance, and an increase of 24% with author assistance for the articles published in *Psychological Science* that received an 'open data badge'. Similar reproducibility rates were reported for political science (Eubank, 2016; Stockemer et al., 2018) and economics (Chang & Li, 2015).

Only two studies assessing the reproducibility of publications in psychology report much higher estimates: Obels et al. (2020) report 58% and Artner et al. (2021) – 70% success rate. We believe that several factors may contribute to the difference in estimates: first, there is likely a sampling bias. Obels et al. assessed the reproducibility of registered

Table 4
Summary of reproducibility rates in published investigations.

Paper	Reproduced / total N	Percentage reproduced (95% CI)
Stodden et al. (2018)	21 (extrapolated to 53) / 204*	26 [20, 32]%
Hardwicke et al. (2018)	11 / 35	31 [17, 51]%
Hardwicke et al. (2021)	9 / 25	36 [20, 59]%
Obels et al. (2020)	21 / 36	58 [39, 72]%
Eubank (2016)	4 / 24	17 [4, 29]%
Stockemer et al. (2018)	32 / 70	46 [33, 56]%
Chang and Li (2015)	22 / 59	37 [24, 49]%
Artner et al. (2021)	163 / 232 (46 articles)	70 [64 , 75]%
Naudet et al. (2018)	14 / 17	82 [53, 94]%

The asterisk (*) marks the fact that 56 articles were deemed reproducible by preliminary inspection. Of these, 22 were tested, and 21 were successfully reproduced. The reproducibility rate of 26% is an extrapolation to the untested articles.

reports (see Nosek & Lakens, 2014), and Artner et al. (2021) – the reproducibility of the papers whose authors shared data upon request. Wicherts et al. (2011) suggest that willingness to share data could be associated with the strength of evidence and the quality of reporting of statistical results. Second, Artner et al. were able to reproduce 70% of target values only after exploring every possible analysis, including those in conflict with the reported analysis procedure, and investing 280 workdays into reproducing 232 reported values. Finally, Obels et al. attempted to reproduce only those papers that shared the analysis code. The authors report that 58% (36 out of 62 papers that shared the data) shared the code, which is close to our estimate of 63%. In our sample, the presence of analysis code increased reproducibility, and the reproducibility rate for papers with code was very close to the one reported by Obels et al.: our estimate was 59% (95% CI [38, 72]%), and the Obels et al. estimate was 58% (95% CI [39, 72]%).[3] Clearly, reproducing only those studies that share analysis code is likely to increase the estimate of reproducibility.

Despite these limitations, we can still make a crude comparison between our reproducibility attempt and those of others in related fields. Reproducibility rates evaluated

[3] Because Obels et al. did not report confidence intervals, we report bootstrapped intervals here for both our and the Obels et al. estimates.

according to the strict criterion are similar in our sample and in the two evaluations of psychological papers by Hardwicke et al. Given that Hardwicke et al. attempted to reproduce only one key finding in each manuscript, and we evaluated all results reported in each manuscript, our lower estimate of 34% success rate is more conservative, and the reproducibility of the surveyed papers is at least as high as that of findings in psychology.

The upper bound for reproducibility in our sample, 56%, is calculated based on the reproducibility criterion that tolerates any number of minor – up to 10% – deviations and up to 20 major discrepancies between the reported and reproduced values. To our knowledge, this is the most lenient criterion of reproducibility of all previously used. And yet, the success rate of 56% is not strikingly high. In almost half of the surveyed papers, the barriers to reproducibility were high enough to render the reported results irreproducible even when evaluated according to such a liberal criterion.

The important questions that arise from our analysis are: What are the major barriers to reproducibility and what can we do as a field to improve reproducibility?

Regarding the first question, we see four main problems: there are issues with the publicly released data, with the code, lack of documentation, and unclear terms of use. These issues are listed below.

1. **Problems with the data**:

    (a) Not all data that is reported in the paper is shared. For 8 papers, data sets were inaccessible. Of 51 papers with accessible data, only 38 (75%) shared all the data necessary to reproduce the reported results. Examples of missing data:

        i. Only aggregated data is shared for a study that analyzed unaggregated data (N=3).

        ii. Five experiments are reported in the paper, but there is data for only three experiments. Or four experiments are reported, but there is data only for three of them, etc. (N=7).

iii. Some predictor variables / covariates used in the analysis are not in the data file (N=3).

iv. The data file that is used in the script is simply absent in the materials (N=6).

v. The data set used in the paper is an openly available corpus, but the novel annotation of the corpus done by the authors was not made available (N=1). Without that annotation, it is impossible to reproduce the analysis.

2. **Problems relating to the analysis code**:

   (a) Not all the code necessary for reproducing the analysis is present (N=5).

   (b) Technical problems with the code:

      i. The code produces an error message that we cannot debug (N=1).

      ii. The code is outdated and does not run any more (this is sometimes referred to as code rot). This happens because the environment in which the code was created changes; e.g., the syntax used for certain functions changes fundamentally, or some needed library is no longer publicly available, etc. (N=1).

      iii. The code is written in proprietary software and cannot be executed free of charge (N=1).

      iv. The modeling code has no documentation or readme; it is unclear what parameter values allow to reproduce the reported results (N=1).

3. **Problems with documentation**:

   (a) The shared data is difficult to interpret:

      i. No readme file clarifies what opaque variable labels stand for. In many cases, it is possible to reconstruct which column names correspond to the variables mentioned in the manuscript. In some cases, it is very hard or impossible.

For three papers, missing mapping from the reported variables to column names was the main obstacle to reproducibility.

ii. The names of the data files do not correspond to how the experiments are numbered or presented in the manuscript, and there is no readme file that establishes the correspondence (N=5).

(b) The analysis is described in insufficient detail or the description contradicts the outcome of the analysis. Examples:

i. The manuscript claims that something was done that is not mirrored in the outcome of the analysis (e.g., the authors claim to have used X as a predictor in the model, but the table summarizing the outcome of the model does not include X as a predictor; the authors claim to have used a particular reference level for a factor variable, but the table summarizing the outcome shows that another factor level was used as a reference, etc.).

ii. The authors state that they report the outcome of a stepwise elimination procedure for model selection, but forget to do so for one of the models, and without this information, the outcome cannot be reproduced.

iii. Data trimming is often described very briefly and ambiguously, so that three different trimming procedures could fit the same description. Or data trimming is not mentioned at all. Often, we could not reproduce the results following the analysis steps described in the manuscript, presumably because the authors did not report some preliminary manipulations with the data set – trimming, centering and scaling of variables, etc. This could be easily avoided if analysis code was present.

4. **Unclear terms of use**: A common problem is that often no license is specified for the data and code (N=33). The absence of a license means that data and code could be under copyright protection and may not be usable by other researchers. It is not

clear to us what the status is of data that have no license or terms of use specified for it. On the one hand, if code is released and no prohibition against further use is stipulated, the code and data should be usable by a third party. On the other hand, data and code are intellectual property, and so should be released with a license. A recent development is that funding agencies have started requiring data management plans or DMPs for experimental research (for an example, see https://osf.io/7c8sz/). DMPs usually require that one specify terms of use (e.g., a waiver) or a license. Ball (2011) is a useful guide to deciding on which license to use.

The surveyed papers exhibit different constellations of the problems we listed. Not all of these problems completely block reproducibility, but all of them hinder it in one way or another. Moreover, for some papers, the precise cause of irreprducibility could not be identified: sometimes the reported values could not be reproduced despite having the full data set and analysis code. In other cases, the data set was published and no trimming procedure was described in the paper, but the condition means could not be reproduced. We tentatively assume that is such cases, the wrong version of either the data set or code was made public, or data and code were not updated after the manuscript had been revised.

We turn now to the question of what steps can be taken to improve reproducibility.

Steps that researchers can take to improve reproducibility

An important point to appreciate is that reproducibility is inherently very difficult to achieve. The reasons for this can be quite mundane, such as a software update. The authors of this manuscript are sometimes unable to reproduce their own analyses, sometimes even a few days after preparing the code. However, with a better workflow, it should be possible to achieve reproducibility. Below, we provide some suggestions for improving the reproducibility of analyses.

1. **Release complete data in a usable form, with preprocessing code**: Just sharing some materials related to the project is not sufficient. Make sure that all the

data necessary to reproduce the analyses is shared. For example, if reporting that a covariate influences a particular conclusion (for example, the position of the trial in the experiment, or participants' age), this variable must be present in the data set.

Share all the data collected: the data file need not consist of the raw output from the recording device; it can have been preprocessed. This is sometimes necessary because otherwise a very large amount of data would have to be stored online. However, all trimming and data transformations (including aggregation) should be performed in a pre-processing code script, and that script should be released with the data. In this context, the "good enough" computing practices advocated by G. Wilson et al. (2017) are worth adopting.

One reason for releasing the raw data is that some attributes of the data set may be valuable to other researchers. For example, the trial ID, the reading times on the last word in the sentence, or the average reading speed may be of no interest in the original paper, but could be highly relevant for researchers who study adaptation effects, sentence wrap-up effects, or average reading speed (Brysbaert, 2019). Public data sets can be reused in ways that the authors haven't thought about (e.g., in meta-analyses, Bürki et al., 2020; Mahowald et al., 2016).

2. **Use publicly accessible repositories for data sharing**: Share data on a publicly accessible server that can guarantee data being available for at least 10 years. Store all the data related to a project in one place: it is easier to maintain and keep up to date a single repository than several copies of the same data set across several repositories. It is generally not a good idea to store data on one's university homepage, because university homepages tend to frequently migrate and change urls. When researchers themselves move from one university to another, their old pages shut down and are no longer available. Providing data and code as part of the journal's supplementary section makes the materials effectively inaccessible to those who do not have a paid

subscription. In contrast to manuscripts, which can be legally accessed as preprints on archives like arXiv or PsyArXiv, or on author's personal websites, these data sets are truly inaccessible for many people.

Most papers that we surveyed shared their data set on Open Science Framework (https://osf.io/), Mendeley data (https://data.mendeley.com/), Zenodo (https://zenodo.org/) and some open institutional repositories. Other existing solutions include the Dataverse Project (https://dataverse.org/) and Github (https://github.com).

3. **Use non-proprietary data formats**: Sometimes data files are released as Word documents, Excel sheets, or even as pdf files. It is better to use plain text or comma-separated-values (csv) files for data release because in contrast to proprietary formats, these formats do not get outdated: it might be impossible to open a word file created some 20 years ago, but plain text files can still be read.

   Storing data in Excel sheets is also problematic because Excel makes some automatic decisions behind the scenes, for example, regarding the type of the input data. Researchers may not even be aware that the representation of their data has changed. This has already led to serious mistakes: for example, several human gene names are automatically converted to dates, which affected up to one-fifth of the surveyed genomics papers (Ziemann et al., 2016) and ultimately led to changes in the guidelines for genome naming (Wain et al., 2002); another example involves failure to report new COVID cases in the UK because an Excel spreadsheet containing lab results reached its maximum size, and failed to update for over a week.[4]

   Using Excel for data analysis is even more dangerous: Excel formulas can be scattered through several sheets and are notoriously hard to test and debug, which leads to the proliferation of errors. An example from economics is discussed in Herndon et al.

[4] https://www.bbc.com/news/technology-54423988

(2014).

4. **Provide documentation**: Write a description of the data set and variable names (usually stored in a file called "readme", "codebook" or "data dictionary"). Don't be afraid of redundancy: In addition to having the codebook, give variables and their values transparent and self-explanatory names. For example, if there are two conditions like subject relatives and object relatives, label the levels "subj-rel" and "obj-rel" or the like, instead of non-obvious labels like 1 and 2. Provide meta-data if the chosen storage server allows it: meta-data makes the data set more findable, which increases the chances that it will be reused and cited.

   Related to documentation, the experiment item files (including fillers) should also be provided. These are usually not needed to ensure reproducibility, but can be very valuable for carrying out replication attempts.

5. **Share analysis code together with data.** According to our estimates, providing code with data increases reproducibility by 38% (95% CrI [19, 60]%). So, even this one minimal change will have a big impact. Conversely, if the analysis code is not shared, it becomes very difficult and time-consuming, if not impossible, to trace back how the reported results were obtained.

   A nice illustration of the cost incurred by non-reproducible code is the case recently reported by Brewer et al. (2021): two analyses of the same data set yielded completely different results, and "it took the authors several days of email correspondence to determine where the differences were coming from" (the main difference was the mean-centering of predictors in one of the analyses). Two studies of reproducibility in psychology report similar estimates: "5–25 person hours" (Hardwicke et al., 2018, p. 12), and "between 1 and 30 (median = 7, interquartile range = 5) hours" (Hardwicke et al., 2021, p. 5) to reproduce one key result from a published paper. Scenarios like these can be avoided by simply providing reproducible analysis code.

A further complication is that the cases described above refer to the lucky scenario when the authors of the original manuscript are responsive and actively participate in the efforts to reproduce their analyses. Often this is not the case (Vanpaemel et al., 2015; Wicherts et al., 2006); in addition, author responsiveness and ability to provide information decrease with time after publication (Minocher et al., 2020; Vines et al., 2014), mostly due to email addresses that no longer work, and to lost or otherwise inaccessible data sets. For all of these reasons, it is highly desirable that all experimental materials, such as the data set, experimental items, and data analysis code, are shared in such a way that they can be used on a stand-alone basis, not dependent on the authors being able and willing to provide clarifications.

6. **Test the code and data release for usability**: After uploading data and code to a server, try downloading it to a different folder, preferably on a different computer, and running the analysis. A common outcome of such testing is that the uploaded analysis is not yet ready for release: some code or data files are not yet in the repository, some of the required software packages are not explicitly loaded or mentioned, etc. This can happen, for example, if the researchers rely on automatically saving some of the objects they created in a hidden .Rdata file, which is not available to a third-party user. After fixing the problems that arose as a result of this test, one may ask a colleague to download one's materials and reproduce the analysis, to test whether a person who is not intimately familiar with the project can understand and run the analysis.

7. **Teach the next generation to use data management and computing skills**: A root cause for the problems that researchers face in producing reproducible research is that in psychology and psycholinguistics, historically there has been no tradition of teaching beginning researchers how to develop a research workflow that includes good quality code writing practices, data management, and documentation techniques. A good place to start would be to host a Data Carpentry (https://datacarpentry.org/) or

a similar workshop on data management.

Following these suggestions should eliminate the majority of the problems that we encountered. There are many recommendations and tutorials available for developing a better workflow. For a set of simple steps that increase reproducibility, refer to Obels et al. (2020) and G. Wilson et al. (2017); for a gentle introduction to data sharing steps, such as choosing a repository, preparing data and code for sharing, structuring folders, see O. Klein et al. (2018) and https://www.ukrn.org/primers/; for the illustration of how the principles of high-reliability organizations, such as aviation and medicine, can be applied in the context of a scientific lab in psychology and related areas, see Rouder et al. (2019); for a discussion of "Findable, Accessible, Interoperable, and Reusable" or FAIR principles of data sharing and their implementation in practice, see Jacobsen et al. (2020); and for a more technically involved set of recommendations for organizing a data analysis workflow, see Peikert and Brandmaier (2021), Peikert et al. (2021).

Some good examples of code and data release that we found during our assessment are the following: (i) Boyce et al. (2020): https://github.com/vboyce/Maze, (ii) Siew et al. (2021): https://osf.io/adc2p/, (iii) Nooteboom and Quené (2020): https://osf.io/jahqe/, and (iv) Günther, Nguyen, et al. (2020): https://osf.io/ftxjy/.

Shared data and code, unlike the text of the manuscript, can be updated even after publication. For that reason, before adopting new data sharing practices in the ongoing or future projects, researchers might want to test their already published materials for usability, and update them accordingly. Such a self-administered reproducibility check is useful for two reasons: for the researcher, it is a very informative first step highlighting their current weak spots and areas of improvement; for the scientific community, it makes the materials truly available. After all, for a researcher accessing materials ten years from now, it would not matter whether initially, some of the materials were missing; all that matters is that the materials are available at the time of access.

One important change that researchers might optionally consider adopting in their

workflow is to use automated ways to incorporate numbers and figures in a manuscript. In our attempts to reproduce the reported results, we encountered copy-and-paste errors, such as two identical rows in a table summarizing the results of a linear model, or the loss of all minus signs in a large table. Copying and pasting numbers by hand is time-consuming and error-prone. Good automated alternatives for R include excellent packages `sjPlot` (Lüdecke, n.d.) and `apaTables` (Stanley & Spence, 2018) producing publication-ready summary tables for Word and html, the `kableExtra` package producing tables for LaTeX(Zhu, 2019), and the `report` package for automatic generation of both tables and texts reporting the outcomes of statistical analyses (Makowski et al., 2021). R Markdown considerably simplifies the manuscript-preparation process; packages like `papaja` (Aust & Barth, 2018) allow the researcher to automatically format the paper in APA style, and to dynamically generate tables, figures, and numbers within the manuscript. Under this literate programming style (Knuth, 1984), the manuscript also serves as a complete documentation of the data analysis workflow. The present paper is written using a literate programming style.

Potential concerns to developing a reproducible workflow

Here, we address some concerns that researchers might have to developing a reproducible workflow.

A concern shared by 22% of psychology researchers (Borghi & Van Gulick, 2021) is that writing code and documentation in a publicly accessible way will be very time-consuming and will take time away from core scientific work. One may also lose time if one has to acquire new data management and coding skills (e.g., learning R Markdown for document preparation).

There is no denying that preparing the data set and analysis code for archiving is indeed costly. Three considerations may help here. First, if the researcher adopts a reproducible workflow, the initial time investment is likely to be fully repaid during the active phase of working on a project. For example, the entire analysis for a slightly changed

data file can be carried out by simply re-running the analysis script. If one adopts a literate programming style, all the tables, figures, and in-text numbers are generated automatically. Second, when preparing the data and code for public release, the researcher is likely to pay more attention to the analysis and may catch errors that may otherwise have gone unnoticed. Catching such errors and ensuring the reliability of published findings is core scientific work, and is worth the time cost. Finally, if, as required by many journals, code and data are released during the review process, mistakes can be caught before the paper is published. For example, in one case (Jäger et al., 2020), a reviewer (Brian Dillon) examined the code and found an error in the code while reviewing the paper; the error was corrected in the revision. If the authors hadn't provided the code and data, this error could have gone undetected.

Another potential concern is: will anyone ever look at the published data and code? If not, why invest time and effort into releasing these? It is indeed impossible to know whether other researchers will engage with one's data, but if the data and code are not shared, it is guaranteed that nobody will use it, because nobody can. The easier a data set is to access publicly, the higher the chance that it will be reused. In addition, as mentioned earlier, (some aspects of) the data set may turn out to be useful for answering new research questions, and for evidence synthesis through meta-analysis (Bürki et al., 2020; Cox et al., 2022; Jäger et al., 2017; Mahowald et al., 2016; Nicenboim et al., 2020; Vasishth et al., 2013). Finally, the authors themselves may wish to revisit the published data. It is therefore useful to treat the published materials as a personal back-up. From this point of view, preparing the data set and the analysis code is a courtesy to the researcher's future self who will not need to spend days looking for data and trying to understand which file contains the relevant information in ten years' time.

Several surveys of data sharing practices show that less experienced researchers are hesitant to share data and analysis code out of fear of public shaming, loss of reputation, etc., if any errors are found (Meteyard & Davies, 2020; Soeharjono & Roche, 2021). It is possible that errors may be found, but this does not necessarily lead to an adverse effect on

one's career (Ebersole et al., 2016). While we cannot speak for all researchers, in our experience, usually those who are good at spotting mistakes are good at it precisely because they found a lot of mistakes in their own code, and therefore tend to be more understanding towards mistakes in other people's code. The reality is that everyone makes mistakes; it is not a personal failure on a researcher's part when a mistake is discovered.

One strategy for overcoming the fear of having an error detected publicly is to ask a trusted colleague for feedback before releasing the data and code; this is no different than asking for feedback on a to-be-submitted manuscript. One can go one step further and organize a code review group within one's lab, similarly to writing groups that exist in many research groups. Finally, while mistakes in the code are traditionally associated with compromising the main theoretical claim of the work, this is by no means the most common outcome. In our attempts to reproduce the published analyses, we repeatedly found effects that supported the main theoretical claim of the manuscript but were not reported by the authors (such as a bigger or significant effect of interest in the model for which it was reported as smaller or non-significant). In other words, publishing and reviewing code can strengthen the theoretical claims of the paper instead of casting doubt on them.

In sum, the practices suggested above do come with a price, but we believe that this is a price worth paying both for the benefit of individual researchers and for the global benefit of the scientific community. Reproducibility is important and worth aiming for because it increases trust in science, and enables scientists to feel more confident about what they learned from previous work. This is a very valuable outcome for everyone, and the extra effort required to approach reproducibility will help all researchers.

Changes that journals can make

Journals are already actively instituting policies that are fostering positive change. Below, we list some further changes that journals can make.

1. Make it obligatory to share analysis code along with the data, unless there are

compelling reasons not to do so. The option to write that "the data will be provided on request" should, as a rule, not be allowed, because authors generally do not release data on request.

2. Stipulate a clear section in the paper in which the link to the materials should always appear. Currently, the links can be found all across the paper text, from the first mention of the data set in the Methods section to Acknowledgements. In several papers the official research data statement was "Data not available / Data will be made available on request", but the link to the data set was in fact found in some place in the paper. One paper had three different links to non-identical versions of the associated data set in three different sections of the paper.

3. Have an editorial assistant check the technical requirements upon submission. Specifically, the assistant should check that:

   (a) the link indeed leads to the data and code;

   (b) a license is specified that allows some kind of reuse.

4. For the data sets published on the journal website along with the papers, provide an obligatory slot for licensing information.

5. JML's submission portal automatically renames files. Changed file names make it difficult to establish which data file corresponds to which experiment. The submission portal should never rename data and code files.

6. At the resubmission stage, prompt the authors to upload updated data and analysis files to their repository of choice. Provide a data-sharing check-box: the authors should confirm that the shared materials and code are updated and match the revised manuscript.

7. Provide in-house assessment of reproducibility (for a discussion and analysis of the efficiency of such measures, see Eubank, 2016; Sakaluk et al., 2014). Many journals,

like *Language* and *Glossa Psycholinguistics*, are planning to have or already have statistical consultants on their editorial board.

8. Engage reviewers who are part of the The Peer Reviewers' Openness (PRO) Initiative (Morey et al., 2016) and who therefore will further ensure that open science practices are implemented.

9. Carry out periodic evaluations like the present one to evaluate whether the policies are working.

Many problems that we have listed block the very beginning of the data analysis pipeline (missing data, missing variable descriptions, unclear data selection and trimming procedures), so it is possible that our suggestions, while helping to solve these problems, would expose other problems later in the pipeline that would call for different solutions and quality control procedures. For example, we only considered the reproducibility of the results under the analysis chosen by the authors of the original manuscript. A more challenging assessment of reproducibility would be to test if the key results hold under different, equally well-motivated analyses (for an example of such assessment, see Breznau et al., 2021; Silberzahn et al., 2018). The results that could pass such a test, and be independently replicated, will also serve to contribute to increased reliability in science.

Conclusion

We assessed data availability for 59 papers published in Journal of Memory and Language before and 59 papers published after a mandatory data sharing policy was adopted. The new policy has increased the rate of data sharing by more than 50%: 20 (34%) papers in the pre-policy group vs. 51 (86%) papers in the post-policy group had downloadable data sets. Of the 59 papers published under the data sharing policy that we investigated we were able to fully reproduce 20 (34%). Even under a very lenient reproducibility criterion that tolerates any number of small – up to 10% – differences and up to 20 large-scale

discrepancies between the reproduced and the reported values, the success rate was only 56%. Two simple steps can significantly increase the reproducibility of published papers: first, the authors should share the analysis code along with the data; and second, they should try to download the shared materials and attempt to reproduce their own analyses using only the shared data and code. Such a simple self-administered reproducibility check, similar to the proofreading of a manuscript, should be a part of normal research workflow.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 317633480, SFB 1287. We are grateful to Dario Paape and Dorothea Pregla for comments on a draft.

References

Ahn, S., & Brown-Schmidt, S. (2020). Retrieval processes and audience design. *Journal of Memory and Language*, *115*, 104149.

Akan, M., Stanley, S. E., & Benjamin, A. S. (2018). Testing enhances memory for context. *Journal of Memory and Language*, *103*, 19–27.

Ambrus, G. G., Vékony, T., Janacsek, K., Trimborn, A. B., Kovács, G., & Nemeth, D. (2020). When less is more: Enhanced statistical learning of non-adjacent dependencies after disruption of bilateral DLPFC. *Journal of Memory and Language*, *114*, 104144.

Anderson, J. E., Aarts, A. A., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., Babel, M., Bahniék, Š., Baranski, E., Barnett-Cowan, M., Et al. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Arnold, J. E., Strangmann, I. M., Hwang, H., Zerkle, S., & Nappa, R. (2018). Linguistic experience affects pronoun interpretation. *Journal of Memory and Language*, *102*, 41–54.

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, *26*(5), 527–546.

Aust, F., & Barth, M. (2018). Papaja: Create APA manuscripts with R Markdown.

Avetisyan, S., Lago, S., & Vasishth, S. (2020). Does case marking affect agreement attraction in comprehension? *Journal of Memory and Language*, *112*, 104087.

Ball, A. (2011). How to license research data. *Edinburgh, UK*, Digital Curation Center.

Bangerter, A., Mayor, E., & Knutsen, D. (2020). Lexical entrainment without conceptual pacts? Revisiting the matching task. *Journal of Memory and Language*, *114*, 104129.

Borghi, J. A., & Van Gulick, A. E. (2021). Data management and sharing: Practices and perceptions of psychology researchers. *PloS one*, *16*(5), e0252047.

Boyce, V., Futrell, R., & Levy, R. P. (2020). Maze Made Easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, *111*, 104082.

Brainerd, C., Bialer, D., & Chang, M. (2020). Norming retrieval processes. *Journal of Memory and Language*, *115*, 104143.

Brainerd, C., Nakamura, K., Chang, M., & Bialer, D. (2019). Super-overdistribution. *Journal of Memory and Language*, *108*, 104027.

Brainerd, C., Nakamura, K., & Murtaza, Y. (2020). Explaining complementarity in false memory. *Journal of Memory and Language*, *112*, 104105.

Brandt, M., Aßfalg, A., Zaiser, A.-K., & Bernstein, D. M. (2020). A computational approach to the revelation effect. *Journal of Memory and Language*, *112*, 104091.

Brewer, G. A., Robey, A., & Unsworth, N. (2021). Discrepant findings on the relation between episodic memory and retrieval practice: The impact of analysis decisions. *Journal of Memory and Language*, *116*, 104185.

Breznau, N., Rinke, E. M., Wuttke, A., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., & Bahnsen, O. (2021). *Observing many researchers using the same data and hypothesis reveals a hidden universe of data analysis* (MetaArXiv cd5j9). Center for Open Science. https://EconPapers.repec.org/RePEc:osf:metaar:cd5j9

Bristol, R., & Rossano, F. (2020). Epistemic trespassing and disagreement. *Journal of Memory and Language*, *110*, 104067.

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, *116*, 104174.

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, *109*, 104047.

Bürki, A., Elbuy, S., Madec, S., & Vasishth, S. (2020). What did we learn from forty years of research on semantic interference? A Bayesian meta-analysis. *Journal of Memory and Language*, *114*, 104125.

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software*, *80*(1), 1–28.

Chambers, C. D. (2020). Verification reports: A new article type at Cortex. *Cortex*, *129*, A1–A3.

Chan, J. C., Manley, K. D., & Ahn, D. (2020). Does retrieval potentiate new learning when retrieval stops but new learning continues? *Journal of Memory and Language*, *115*, 104150.

Chan, J. C., Manley, K. D., Davis, S. D., & Szpunar, K. K. (2018). Testing potentiates new learning across a retention interval and a lag: A strategy change perspective. *Journal of Memory and Language*, *102*, 83–96.

Chang, A. C., & Li, P. (2015). Is economics research replicable? Sixty published papers from thirteen journals say 'usually not'.

Chetail, F. (2020). Are graphemic effects real in skilled visual word recognition? *Journal of Memory and Language*, *111*, 104085.

Chubala, C., Surprenant, A. M., Neath, I., & Quinlan, P. T. (2018). Does dynamic visual noise eliminate the concreteness effect in working memory? *Journal of Memory and Language*, *102*, 97–114.

Collins, R. N., Milliken, B., & Jamieson, R. K. (2020). Minerva-de: An instance model of the deficient processing theory. *Journal of Memory and Language*, *115*, 104151.

Corps, R. E., & Rabagliati, H. (2020). How top-down processing enhances comprehension of noise-vocoded speech: Predictions about meaning are more important than predictions about form. *Journal of Memory and Language*, *113*, 104114.

Cox, C. M. M., Keren-Portnoy, T., Roepstorff, A., & Fusaroli, R. (2022). A Bayesian meta-analysis of infants' ability to perceive audio–visual congruence for speech. *Infancy*, *27*(1), 67–96.

Cunnings, I., & Sturt, P. (2018). Retrieval interference and semantic interpretation. *Journal of Memory and Language*, *102*, 16–27.

de Bruin, A., Samuel, A. G., & Duñabeitia, J. A. (2018). Voluntary language switching: When and why do bilinguals switch between their languages? *Journal of Memory and Language*, *103*, 28–43.

Deliens, G., Antoniou, K., Clin, E., Ostashchenko, E., & Kissine, M. (2018). Context, facial expression and prosody in irony processing. *Journal of memory and language*, *99*, 35–48.

Dıéez-Álamo, A. M., Glenberg, A. M., Dıéez, E., Alonso, M. A., & Fernandez, A. (2020). The linguistic looming effect. *Journal of Memory and Language*, *114*, 104147.

Do, M. L., & Kaiser, E. (2019). Subjecthood and linear order in linguistic encoding: Evidence from the real-time production of wh-questions in English and Mandarin Chinese. *Journal of Memory and Language*, *105*, 60–75.

Drummer, J.-D., & Felser, C. (2018). Cataphoric pronoun resolution in native and non-native sentence comprehension. *Journal of Memory and Language*, *101*, 97–113.

Ebersole, C. R., Axt, J. R., & Nosek, B. A. (2016). Scientists' reputations are based on getting it right, not being right. *PLoS biology*, *14*(5), e1002460.

Eubank, N. (2016). Lessons from a decade of replications at the Quarterly Journal of Political Science. *PS: Political Science & Politics*, *49*(2), 273–276.

Falandays, J. B., Brown-Schmidt, S., & Toscano, J. C. (2020). Long-lasting gradient activation of referents during spoken language processing. *Journal of Memory and Language*, *112*, 104088.

Fellman, D., Jylkkä, J., Waris, O., Soveri, A., Ritakallio, L., Haga, S., Salmi, J., Nyman, T. J., & Laine, M. (2020). The role of strategy use in working memory training outcomes. *Journal of Memory and Language*, *110*, 104064.

Fisher, J. S., & Radvansky, G. A. (2018). Patterns of forgetting. *Journal of Memory and Language*, *102*, 130–141.

Floccia, C., Delle Luche, C., Lepadatu, I., Chow, J., Ratnage, P., & Plunkett, K. (2020). Translation equivalent and cross-language semantic priming in bilingual toddlers. *Journal of Memory and Language*, *112*, 104086.

Fox, J., Dennis, S., & Osth, A. F. (2020). Accounting for the build-up of proactive interference across lists in a list length paradigm reveals a dominance of item-noise in recognition memory. *Journal of Memory and Language*, *110*, 104065.

Frazier, L., & Clifton Jr, C. (2018). Topic situations: Coherence by inclusion. *Journal of Memory and Language*, *103*, 176–190.

Fritz, I., Kita, S., Littlemore, J., & Krott, A. (2019). Information packaging in speech shapes information packaging in gesture: The role of speech planning units in the coordination of speech-gesture production. *Journal of Memory and Language*, *104*, 56–69.

Fujita, H., & Cunnings, I. (2020). Reanalysis and lingering misinterpretation of linguistic dependencies in native and non-native sentence comprehension. *Journal of Memory and Language*, *115*, 104154.

Fukumura, K. (2018). Ordering adjectives in referential communication. *Journal of Memory and Language*, *101*, 37–50.

Gagné, C. L., Spalding, T. L., Spicer, P., Wong, D., Rubio, B., & Cruz, K. P. (2020). Is buttercup a kind of cup? Hyponymy and semantic transparency in compound words. *Journal of Memory and Language*, *113*, 104110.

Galati, A., Dale, R., & Duran, N. D. (2019). Social and configural effects on the cognitive dynamics of perspective-taking. *Journal of Memory and Language*, *104*, 1–24.

Garnham, A., Child, S., & Hutton, S. (2020). Anticipating causes and consequences. *Journal of Memory and Language*, *114*, 104130.

Gathercole, S. E., Dunning, D. L., Holmes, J., & Norris, D. (2019). Working memory training involves learning new skills. *Journal of Memory and Language*, *105*, 19–42.

Gelman, A., & Carlin, J. B. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Gerrig, R., & Rastle, K. (2019). New initiatives to promote open science at the Journal of Memory and Language. *Journal of Memory and Language*, *104*, 126–127. https://doi.org/https://doi.org/10.1016/j.jml.2018.10.004

Günther, F., Nguyen, T., Chen, L., Dudschig, C., Kaup, B., & Glenberg, A. M. (2020). Immediate sensorimotor grounding of novel concepts learned from language alone. *Journal of Memory and Language*, *115*, 104172.

Günther, F., Petilli, M. A., & Marelli, M. (2020). Semantic transparency is not invisibility: A computational model of perceptually-grounded conceptual combination in word processing. *Journal of Memory and Language*, *112*, 104104.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Et al. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573.

Hardwicke, T. E., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M. B., Peloquin, B. N., DeMayo, B. E., Long, B., Yoon, E. J., & Frank, M. C. (2021). Analytic reproducibility in articles receiving open data badges at the journal Psychological Science: An observational study. *Royal Society open science*, *8*(1), 201494.

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., Et al. (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.

Healey, M. K. (2018). Temporal contiguity in incidentally encoded memories. *Journal of Memory and Language*, *102*, 28–40.

Herndon, T., Ash, M., & Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, *38*, 257–279.

Hesse, C., & Benz, A. (2020). Scalar bounds and expected values of comparatively modified numerals. *Journal of Memory and Language*, *111*, 104068.

Hollis, G. (2020). Delineating linguistic contexts, and the validity of context diversity as a measure of a word's contextual variability. *Journal of Memory and Language*, *114*, 104146.

Hopper, W. J., & Huber, D. E. (2018). Learning to recall: Examining recall latencies to test an intra-item learning theory of testing effects. *Journal of Memory and Language*, *102*, 1–15.

Hsiao, Y., & Nation, K. (2018). Semantic diversity, frequency and the development of lexical quality in children's word reading. *Journal of Memory and Language*, *103*, 114–126.

Humphreys, M. S., Li, Y. R., Burt, J. S., & Loft, S. (2020). How semantic processing affects recognition memory. *Journal of Memory and Language*, *113*, 104109.

Hwang, H., & Shin, J.-A. (2019). Cumulative effects of syntactic experience in a between-and a within-language context: Evidence for implicit learning. *Journal of Memory and Language*, *109*, 104054.

Isarida, T., Isarida, T. K., Kubota, T., Higuma, M., & Matsuda, Y. (2018). Influences of context load and sensibleness of background photographs on local environmental context-dependent recognition. *Journal of Memory and Language*, *101*, 114–123.

Isarida, T., Isarida, T. K., Kubota, T., Nakajima, S., Yagi, K., Yamamoto, A., & Higuma, M. (2020). Video context-dependent effects in recognition memory. *Journal of Memory and Language*, *113*, 104113.

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W., Imming, M., Jeffery, K. G., ... Schultes, E. (2020). FAIR Principles: Interpretations and implementation considerations. *Data Intelligence*, *2*(1-2), https://direct.mit.edu/dint/article-pdf/2/1-2/10/1893430/dint_r_00024.pdf, 10–29. https://doi.org/10.1162/dint_r_00024

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316–339. https://doi.org/https://doi.org/10.1016/j.jml.2017.01.004

Jäger, L. A., Mertzen, D., Van Dyke, J. A., & Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*, *111*, 104063.

James, A. N., Fraundorf, S. H., Lee, E.-K., & Watson, D. G. (2018). Individual differences in syntactic processing: Is there evidence for reader-text interactions? *Journal of memory and language*, *102*, 155–181.

Johns, B. T., Jamieson, R. K., Crump, M. J., Jones, M. N., & Mewhort, D. (2020). Production without rules: Using an instance memory model to exploit structure in natural language. *Journal of Memory and Language*, *115*, 104165.

Jones, T., & Farrell, S. (2018). Does syntax bias serial order reconstruction of verbal short-term memory? *Journal of Memory and Language*, *100*, 98–122.

Jou, J., Escamilla, E. E., Torres, A. U., Ortiz Jr, A., & Salazar, P. (2018). Where does the congruity effect come from in memorial comparative judgments? A

serial-position-based distinctiveness account. *Journal of Memory and Language*, *103*, 127–150.

Karimi, H., Swaab, T. Y., & Ferreira, F. (2018). Electrophysiological evidence for an independent effect of memory retrieval on referential processing. *Journal of Memory and Language*, *102*, 68–82.

Kaula, A. J., & Henson, R. N. (2020). Priming effects on subsequent episodic memory: Testing attentional accounts. *Journal of Memory and Language*, *113*, 104106.

Keung, L.-C., & Staub, A. (2018). Variable agreement with coordinate subjects is not a form of agreement attraction. *Journal of Memory and Language*, *103*, 1–18.

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

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahnık, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Et al. (2014). Investigating variation in replicability. *Social Psychology*, *45*(3), 142–152.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahnık, Š., Et al. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490.

Knuth, D. E. (1984). Literate programming. *The computer journal*, *27*(2), 97–111.

Kowialiewski, B., & Majerus, S. (2018). The non-strategic nature of linguistic long-term memory effects in verbal short-term memory. *Journal of Memory and Language*, *101*, 64–83.

Lange, N., Berry, C. J., & Hollins, T. J. (2019). Linking repetition priming, recognition, and source memory: A single-system signal-detection account. *Journal of Memory and Language*, *109*, 104039.

Lauro, J., Schwartz, A. I., & Francis, W. S. (2020). Bilingual novel word learning in sentence contexts: Effects of semantic and language variation. *Journal of Memory and Language*, *113*, 104123.

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.

Lelonkiewicz, J. R., Ktori, M., & Crepaldi, D. (2020). Morphemes as letter chunks: Discovering affixes through visual regularities. *Journal of Memory and language*, *115*, 104152.

Li, X., Ren, G., Zheng, Y., & Chen, Y. (2020). How does dialectal experience modulate anticipatory speech processing? *Journal of Memory and Language*, *115*, 104169.

Liang, F., Ma, J., Bai, X., & Liversedge, S. P. (2021). Initial landing position effects on Chinese word learning in children and adults. *Journal of Memory and Language*, *116*, 104183.

Lüdecke, D. (n.d.). *SjPlot: Data visualization for statistics in social science* [R package version 2.8.9]. R package version 2.8.9. https://CRAN.R-project.org/package=sjPlot

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), 1–18.

Mahowald, K., James, A., Futrell, R., & Gibson, E. (2016). A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*, *91*, 5–27.

Makowski, D., Ben-Shachar, M. S., Patil, I., & Lüdecke, D. (2021). Automated results reporting as a practical tool to improve reproducibility and methodological best practices adoption. *CRAN*. https://github.com/easystats/report

Malejka, S., & Bröder, A. (2019). Exploring the shape of signal-detection distributions in individual recognition ROC data. *Journal of Memory and Language*, *104*, 83–107.

McKinley, G. L., & Benjamin, A. S. (2020). The role of retrieval during study: Evidence of reminding from overt rehearsal. *Journal of Memory and Language*, *114*, 104128.

McKoon, G., & Ratcliff, R. (2018). Adults with poor reading skills, older adults, and college students: The meanings they understand during reading using a diffusion model analysis. *Journal of memory and language*, *102*, 115–129.

Mertzen, D., Laurinavichyute, A., Dillon, B., Engbert, R., & Vasishth, S. (2020). A cross-linguistic investigation of proactive, similarity-based retrieval interference in sentence comprehension: No support from English, German and Russian eye-tracking data.

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092.

Miller, A. L., Gross, M. P., & Unsworth, N. (2019). Individual differences in working memory capacity and long-term memory: The influence of intensity of attention to items at encoding as measured by pupil dilation. *Journal of Memory and Language*, *104*, 25–42.

Minocher, R., Atmaca, S., Bavero, C., McElreath, R., & Beheim, B. (2020). Reproducibility improves exponentially over 63 years of social learning research.

Miyoshi, K., Kuwahara, A., & Kawaguchi, J. (2018). Comparing the confidence calculation rules for forced-choice recognition memory: A winner-takes-all rule wins. *Journal of Memory and Language*, *102*, 142–154.

Mohanty, P. P., & Naveh-Benjamin, M. (2018). Mitigating the adverse effects of response deadline on recognition memory: Differential effects of semantic memory support on item and associative memory. *Journal of Memory and Language*, *102*, 182–194.

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

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9.

National Academies of Sciences, E., & Medicine. (2019). *Reproducibility and replicability in science*. Washington, DC, The National Academies Press. https://doi.org/10.17226/25303

Naudet, F., Sakarovitch, C., Janiaud, P., Cristea, I., Fanelli, D., Moher, D., & Ioannidis, J. P. (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. *Bmj*, *360*.

Nicenboim, B., Schad, D. J., & Vasishth, S. (2021). *Introduction to Bayesian data analysis for cognitive science* [Under contract with Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences Series]. https://vasishth.github.io/bayescogsci/

Nicenboim, B., & Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language*, *99*, 1–34.

Nicenboim, B., Vasishth, S., & Rösler, F. (2020). Are words pre-activated probabilistically during sentence comprehension? evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia*, *142*. https://doi.org/10.1016/j.neuropsychologia.2020.107427

Nieuwland, M. S., Arkhipova, Y., & Rodrıéguez-Gómez, P. (2020). Anticipating words during spoken discourse comprehension: A large-scale, pre-registered replication study using brain potentials. *Cortex*, *133*, 1–36.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Zu Wolfsthurn, S. V. G., Bartolozzi, F., Kogan, V., Ito, A., Et al. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468.

Nooteboom, S. G., & Quené, H. (2019). Temporal aspects of self-monitoring for speech errors. *Journal of Memory and Language*, *105*, 43–59.

Nooteboom, S. G., & Quené, H. (2020). Repairing speech errors: Competition as a source of repairs. *Journal of Memory and Language*, *111*, 104069.

Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*, *45*, 137–141.

Nuijten, M. B., Bakker, M., Maassen, E., & Wicherts, J. (2018). Verify original results through reanalysis before replicating. *The Behavioral and Brain Sciences*, *41*, e143–e143. https://doi.org/10.1017/s0140525x18000791

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

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.

Obels, P., Lakens, D., Coles, N. A., Gottfried, J., & Green, S. A. (2020). Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 229–237.

Osth, A. F., Fox, J., McKague, M., Heathcote, A., & Dennis, S. (2018). The list strength effect in source memory: Data and a global matching model. *Journal of Memory and Language*, *103*, 91–113.

Osth, A. F., Shabahang, K. D., Mewhort, D. J., & Heathcote, A. (2020). Global semantic similarity effects in recognition memory: Insights from BEAGLE representations and the diffusion decision model. *Journal of Memory and Language*, *111*, 104071.

Paap, K. R., Anders-Jefferson, R., Mikulinsky, R., Masuda, S., & Mason, L. (2019). On the encapsulation of bilingual language control. *Journal of Memory and Language*, *105*, 76–92.

Paape, D., & Vasishth, S. (2022). *When nothing goes right, go left: An investigation of revisionary and confirmatory rereading using bidirectional self-paced reading* [Submitted]. Submitted.

Peikert, A., & Brandmaier, A. M. (2021). A reproducible data analysis workflow with R Markdown, Git, Make, and Docker. *Quantitative and Computational Methods in Behavioral Sciences*, *1*.

Peikert, A., Van Lissa, C. J., & Brandmaier, A. M. (2021). Reproducible research in R: A tutorial on how to do the same thing more than once. *Psych*, *3*(4), 836–867.

Raff, E. (2019). A step toward quantifying independently reproducible machine learning research. *Advances in Neural Information Processing Systems*, *32*, 5485–5495.

Reifegerste, J., Jarvis, R., & Felser, C. (2020). Effects of chronological age on native and nonnative sentence processing: Evidence from subject-verb agreement in German. *Journal of Memory and Language*, *111*, 104083.

Rouder, J. N., Haaf, J. M., & Snyder, H. K. (2019). Minimizing mistakes in psychological science. *Advances in Methods and Practices in Psychological Science*, *2*(1), 3–11.

Sahakyan, L., & Malmberg, K. J. (2018). Divided attention during encoding causes separate memory traces to be encoded for repeated events. *Journal of Memory and Language*, *101*, 153–161.

Saito, K., Kachlicka, M., Sun, H., & Tierney, A. (2020). Domain-general auditory processing as an anchor of post-pubertal second language pronunciation learning: Behavioural and neurophysiological investigations of perceptual acuity, age, experience, development, and attainment. *Journal of Memory and Language*, *115*, 104168.

Sakaluk, J., Williams, A., & Biernat, M. (2014). Analytic review as a solution to the misreporting of statistical results in psychological science. *Perspectives on Psychological Science*, *9*(6), 652–660.

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, *111*, 104070.

Schubert, T. M., Cohen, T., & Fischer-Baum, S. (2020). Reading the written language environment: Learning orthographic structure from statistical regularities. *Journal of Memory and Language*, *114*, 104148.

Scott, N. M., & Sera, M. D. (2018). Language unifies relational coding: The roles of label acquisition and accessibility in making flexible relational judgments. *Journal of memory and language*, *101*, 136–152.

Seabrooke, T., Hollins, T. J., Kent, C., Wills, A. J., & Mitchell, C. J. (2019). Learning from failure: Errorful generation improves memory for items, not associations. *Journal of Memory and Language*, *104*, 70–82.

Seedorff, M., Oleson, J., & McMurray, B. (2018). Detecting when timeseries differ: Using the Bootstrapped Differences of Timeseries (BDOTS) to analyze Visual World Paradigm data (and more). *Journal of memory and language*, *102*, 55–67.

Siegelman, N., Rueckl, J. G., Steacy, L. M., Frost, S. J., van den Bunt, M., Zevin, J. D., Seidenberg, M. S., Pugh, K. R., Compton, D. L., & Morris, R. D. (2020). Individual differences in learning the regularities between orthography, phonology and semantics predict early reading skills. *Journal of Memory and Language*, *114*, 104145.

Siew, C. S., Yi, K., & Lee, C. H. (2021). Syllable and letter similarity effects in Korean: Insights from the Korean Lexicon Project. *Journal of Memory and Language*, *116*, 104170.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahnıék, Š., Bai, F., Bannard, C., Bonnier, E., Et al. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356.

Singh, K. A., Gignac, G. E., Brydges, C. R., & Ecker, U. K. (2018). Working memory capacity mediates the relationship between removal and fluid intelligence. *Journal of Memory and Language*, *101*, 18–36.

Skrzypulec, B., & Chuderski, A. (2020). Nonlinear effects of spatial connectedness implicate hierarchically structured representations in visual working memory. *Journal of Memory and Language*, *113*, 104124.

Slioussar, N. (2018). Forms and features: The role of syncretism in number agreement attraction. *Journal of Memory and Language*, *101*, 51–63.

Snefjella, B., Lana, N., & Kuperman, V. (2020). How emotion is learned: Semantic learning of novel words in emotional contexts. *Journal of Memory and Language*, *115*, 104171.

Snell, J., & Theeuwes, J. (2020). A story about statistical learning in a story: Regularities impact eye movements during book reading. *Journal of Memory and Language*, *113*, 104127.

Soeharjono, S., & Roche, D. G. (2021). Reported individual costs and benefits of sharing open data among Canadian academic faculty in ecology and evolution. *BioScience*, *71*(7), 750–756.

Stack, C. M. H., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*, *46*(6), 864–877.

Stagge, J. H., Rosenberg, D. E., Abdallah, A. M., Akbar, H., Attallah, N. A., & James, R. (2019). Assessing data availability and research reproducibility in hydrology and water resources. *Scientific Data*, *6*(1), 1–12.

Stanley, D. J., & Spence, J. R. (2018). Reproducible tables in psychology using the apaTables package. *Advances in Methods and Practices in Psychological Science*, *1*(3), 415–431.

Stefanidi, A., Ellis, D. M., & Brewer, G. A. (2018). Free recall dynamics in value-directed remembering. *Journal of Memory and Language*, *100*, 18–31.

Stockemer, D., Koehler, S., & Lentz, T. (2018). Data access, transparency, and replication: New insights from the political behavior literature. *PS: Political Science & Politics*, *51*(4), 799–803.

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584–2589.

Susser, J. A., & Mulligan, N. W. (2019). Exploring the intrinsic-extrinsic distinction in prospective metamemory. *Journal of Memory and Language*, *104*, 43–55.

Thalmann, M., Souza, A. S., & Oberauer, K. (2019). Revisiting the attentional demands of rehearsal in working-memory tasks. *Journal of Memory and Language*, *105*, 1–18.

Tirso, R., & Geraci, L. (2020). Taking another perspective on overconfidence in cognitive ability: A comparison of self and other metacognitive judgments. *Journal of Memory and Language*, *114*, 104132.

Towse, J. N., Ellis, D. A., & Towse, A. S. (2021). Opening Pandora's Box: Peeking inside Psychology's data sharing practices, and seven recommendations for change. *Behavior Research Methods*, *53*(4), 1455–1468.

Troyer, M., & Kutas, M. (2020). To catch a Snitch: Brain potentials reveal variability in the functional organization of (fictional) world knowledge during reading. *Journal of Memory and Language*, *113*, 104111.

Tsuboi, N., & Francis, W. S. (2020). Rethinking bilingual enhancement effects in associative learning of foreign language vocabulary: The role of proficiency in the mediating language. *Journal of Memory and Language*, *115*, 104155.

Uner, O., & Roediger III, H. L. (2018). Are encoding/retrieval interactions in recall driven by remembering, knowing, or both? *Journal of Memory and Language*, *103*, 44–57.

Van Bergen, G., & Bosker, H. R. (2018). Linguistic expectation management in online discourse processing: An investigation of Dutch inderdaad 'indeed' and eigenlijk 'actually'. *Journal of Memory and Language*, *103*, 191–209.

van Heugten, M., Paquette-Smith, M., Krieger, D. R., & Johnson, E. K. (2018). Infants' recognition of foreign-accented words: Flexible yet precise signal-to-word mapping strategies. *Journal of Memory and Language*, *100*, 51–60.

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*, *1*(1).

van Tiel, B., Pankratz, E., & Sun, C. (2019). Scales and scalarity: Processing scalar inferences. *Journal of Memory and Language*, *105*, 93–107.

Vasishth, S., Chen, Z., Li, Q., & Guo, G. (2013). Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PLoS ONE*, *8*(10), 1–14.

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, *59*, 1311–1342. https://doi.org/10.1515/ling-2019-0051

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175.

Vaughn, C., & Kendall, T. (2018). Listener sensitivity to probabilistic conditioning of sociolinguistic variables: The case of (ING). *Journal of Memory and Language*, *103*, 58–73.

Veldre, A., & Andrews, S. (2018a). Beyond cloze probability: Parafoveal processing of semantic and syntactic information during reading. *Journal of Memory and Language*, *100*, 1–17.

Veldre, A., & Andrews, S. (2018b). How does foveal processing difficulty affect parafoveal processing during reading? *Journal of Memory and Language*, *103*, 74–90.

Villani, C., Lugli, L., Liuzza, M. T., Nicoletti, R., & Borghi, A. M. (2021). Sensorimotor and interoceptive dimensions in concrete and abstract concepts. *Journal of Memory and Language*, *116*, 104173.

Vines, T. H., Albert, A. Y., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert, K. J., Moore, J.-S., Renaut, S., & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current Biology*, *24*(1), 94–97.

Wain, H. M., Bruford, E. A., Lovering, R. C., Lush, M. J., Wright, M. W., & Povey, S. (2002). Guidelines for human gene nomenclature. *Genomics*, *79*(4), 464–470.

Wang, J., Otgaar, H., Howe, M. L., Lippe, F., & Smeets, T. (2018). The nature and consequences of false memories for visual stimuli. *Journal of Memory and Language*, *101*, 124–135.

Wedel, A., Nelson, N., & Sharp, R. (2018). The phonetic specificity of contrastive hyperarticulation in natural speech. *Journal of Memory and Language*, *100*, 61–88.

Wen, Y., & van Heuven, W. J. (2018). Limitations of translation activation in masked priming: Behavioural evidence from Chinese-English bilinguals and computational modelling. *Journal of Memory and Language*, *101*, 84–96.

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., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726.

Wilson, B. M., Donnelly, K., Christenfeld, N., & Wixted, J. T. (2019). Making sense of sequential lineups: An experimental and theoretical analysis of position effects. *Journal of Memory and Language*, *104*, 108–125.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. *PLoS Computational Biology*, *13*(6), e1005510.

Yang, H., Hino, Y., Chen, J., Yoshihara, M., Nakayama, M., Xue, J., & Lupker, S. J. (2020). The origins of backward priming effects in logographic scripts for four-character words. *Journal of Memory and Language*, *113*, 104107.

Yim, H., Osth, A. F., Sloutsky, V. M., & Dennis, S. J. (2018). Evidence for the use of three-way binding structures in associative and source recognition. *Journal of Memory and Language*, *100*, 89–97.

Zawadzka, K., Simkiss, N., & Hanczakowski, M. (2018). Remind me of the context: Memory and metacognition at restudy. *Journal of Memory and Language*, *101*, 1–17.

Zhang, X., & Samuel, A. G. (2018). Is speech recognition automatic? Lexical competition, but not initial lexical access, requires cognitive resources. *Journal of Memory and Language*, *100*, 32–50.

Zhu, H. (2019). KableExtra: Construct complex table with 'kable' and pipe syntax. *R package version*, *1*(0).

Ziemann, M., Eren, Y., & El-Osta, A. (2016). Gene name errors are widespread in the scientific literature. *Genome Biology*, *17*(1), 1–3.