Reproducibility of social learning research declines exponentially over 63 years of publication

Riana Minocher[1], Silke Atmaca[1], Claudia Bavero[1], Richard McElreath[1], and Bret Beheim[1]

[1]Department of Human Behaviour, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

June 23, 2020

Abstract

Interest in improving reproducibility, replicability and transparency of research has increased substantially across scientific fields over the last few decades. We surveyed 560 empirical, quantitative publications published between 1955 and 2018, to estimate the rate of reproducibility for research on social learning, a large subfield of behavioural ecology. We found supporting materials were available for less than 30% of publications during this period. The availability of data declines exponentially with time since publication, with a half-life of about six years, and this "data decay rate" varies systematically with both study design and study species. Conditional on materials being available, we estimate that a reasonable researcher could expect to successfully reproduce about 80% of published results, based on our evaluating a subset of 40 publications. Taken together, this indicates an overall success rate of 24% for both acquiring materials and recovering published results, with non-reproducibility of results primarily due to unavailable, incomplete, or poorly-documented data. We provide recommendations to improve the reproducibility of research on the ecology and evolution of social behaviour.

1 Introduction

It is now known that across scientific disciplines, including biology (Andrew et al., 2015), epidemiology (Begley and Ioannidis, 2015), genetics (Ioannidis et al., 2009), and economics (Camerer et al., 2016; Christensen and Miguel, 2018; Chang and Li, 2015), scientific results often do not reproduce or replicate. In the behavioural sciences in particular, the "replication crisis" of psychology (Pashler and Wagenmakers, 2012) has spurred much research on the causes and consequences of non-replication and irreproducibility (OSF, 2015; Maxwell et al., 2015; Klein et al., 2018; Muthukrishna and Henrich, 2019; Smaldino and McElreath, 2016; Nissen et al., 2016; Higginson and Munafa, 2016). This has led to the development and implementation of numerous policies to improve research reliability, such as preregistration to more-clearly distinguish confirmatory and exploratory modes of analysis (Nosek et al., 2018), mandatory data sharing policies (Hardwicke et al., 2018), and improved community guidelines on research design, reporting and data management (Munafò et al., 2017; Ioannidis et al., 2014; McNutt, 2014; Wilkinson et al., 2016) .

Replication, or re-establishing results with new data (Goodman et al., 2016), may be critical to confirm the outcome of a proposed treatment or intervention, for example during clinical trials. However, it remains debatable whether replication should be a goal for all scientific research (Plesser, 2018; Devezer et al., 2019). Replicability does not indicate the quality of research, and not all results may be expected to replicate. For long-lived species such as humans, there are huge costs to collecting long-term data over multiple generations, and it is impossible to engineer a repeated occurrence of a temporally specific phenomenon. A vast array of factors, from demography, local environments, historical contingency and possibly observer bias may thus make it theoretically irrelevant and methodologically infeasible to replicate the patterns that are often a focus of evolutionary studies of behaviour (Schnitzer and Carson, 2016).

While all research need not be replicable, we argue that it should be reproducible. This means an independent researcher can re-establish findings using the data and methods from the original research, with a reasonable amount of effort, as defined by a shared scholarly community (Goodman et al., 2016; Peng, 2011). Reproducibility entails commitment to sharing data either publicly or "upon request", along with unambiguous data processing and analysis details. Reproducibility ensures credibility of published work, improves the cogency of published analyses, and possibly increases productivity and research recognition by streamlining workflows and increasing efficiency (Piwowar et al., 2007), thus promoting cumulative research literatures.

Despite being recognised as a basic responsibility (Baker, 2016; Huang et al., 2012), reproducibility is difficult to achieve. Research reproducibility declines substantially with time since publication (Vines et al., 2014), and fails even when policies to enforce reproducible norms are implemented (Stodden et al., 2018; Baker et al., 2014). The state of research reproducibility within human evolutionary science, specifically, has not been previously assessed. Previous surveys of reproducibility in related fields, such as biology, psychology and ecology, have focused on a single type of analysis (Gilbert et al., 2012), were limited to assessing the implementation of a single journal's policy (Hardwicke et al., 2018), or evaluated only partial aspects of reproducibility, such as

availability of data (Vines et al., 2014). Being methodologically and topically diverse, research in our field offers the potential to investigate the implications of specific barriers to reproducibility, such as variation in research traditions within subfields, or varying constraints on sharing different types of data.

To this end, we surveyed the state of reproducibility within a key literature relevant to evolutionary studies of human behaviour, the study of social learning. We aggregated empirical publications on this topic, spanning a wide range of study species, academic disciplines and research designs. We assessed the availability of data within our sample, and, conditional on its availability, we chose a subset of papers to evaluate the probability of recovering published results. In doing this, we consider how availability of data varies over time and publication type, as well as the causes of non-reproducibility in reanalysis, to identify practices that could better facilitate data sharing and reuse in our field.

2 Methods

2.1 Sampling protocol and description

We sampled empirical, quantitative research relevant to the topic of social learning, including both experimental and observational work, with human and non-human study subjects (Supplementary Information). Our final sample included 560 empirical, quantitative papers published between 1955 and 2018. This comprised 446 experimental and 114 observational studies; 194 studies that included human subjects and 366 studies of non-human animals (non-human primates, birds, reptiles, fish, and other small-bodied mammals). We identified 957 unique authors; each appeared on the author list of up to 49 papers in our sample, with a median number of one paper per author.

We attempted to access materials to reproduce the results of the 560 publications included in our sample, by searching online and sending direct data requests to authors (Supplementary Information). At the end of our sampling and request period, we categorised each study as one of: 'material online'; 'material available through direct request'; 'material not easily accessible to author and/or requesting researcher'; 'material lost'; 'no information on material, i.e. no data statements, no reply to request, or no contact details'.

2.2 Estimating reproducibility

We focus on two aspects of research reproducibility: "empirical reproducibility", meaning the availability of data and other materials to aid reconstruction of analyses, and "analytical reproducibility", meaning the usability of data and other materials to recreate published findings.

To estimate the empirical reproducibility of social learning work, we modelled the probability of obtaining materials for any given study, either publicly accessible or received through a direct data request, with an exponential decay function of year of

publication (Supplementary Information), a standard time-to-event survival model. This allows estimation of the "data decay rate", or percentage decline in availability over time, of social learning data. To assess whether the type of data included in a study affects the rate of decay, we further estimated partially-pooled decay rates for each type of study (experimental and observational data types with human and non-human study subjects). We expected that there could be different constraints on sharing data that includes human subjects, due to privacy and ethical regulations. Concomitantly, experimental data may be easier than observational data to share, having been generated in a specific format and unlikely to be re-used or re-analysed by the experimenter, in contrast to observational data, which is often used repeatedly, to investigate different research questions.

To estimate the analytical reproducibility of results, we selected a random subset of 40 studies from the total sample of studies with materials available (Supplementary Information). For each paper in the subset, we identified individually "citable" results from the abstract of the publication. We located corresponding in-text references for each result, to establish evidence for each in the form of figures, tables or estimates. For a single paper that did not contain an abstract, we identified the main results from the results section of the paper. Using the materials available to us, we then attempted to independently verify the results of each paper. We coded results as "reproduced" if we were able to recreate at least one line of evidence per result, allowing for non-exact numerical reproductions due to differences in software or estimation algorithms. When we were not able to recover a result, we identified the cause of failure at one of three possible states. If we were not able to use the data provided because it was incomplete, such as being too cryptic, raw, or insufficiently annotated, we assessed the cause of failure to be 'data unclear'. If the data were usable, but the analysis was unclear, for example being underdescribed in the text, or too complicated or novel to implement without analytical code or substantial support, we coded the result failure as 'analysis unclear'. When we were able to use the data and reconstruct an analysis as described, if the reconstructed result differed from the stated result in the abstract or text, such as reversing the direction of the stated effect, we coded the cause of failure to be 'reanalysis result differs'. For an example of our reproduction protocol, see https://github.com/rianaminocher/reproducibility-example.

We modelled the probability that a single result of a paper is reproducible, given materials are available, with a logistic link function, thus estimating a global mean probability of reproducibility for all papers sampled. We included partially-pooled intercepts for each paper a result was drawn from, to account for different numbers of results per paper, and to incorporate variation in the probability of reproduction for each paper (Supplementary Information).

From the definition of conditional probability, we calculated a combined probability of empirical and analytical reproducibility of social learning research by multiplying our estimate of empirical reproducibility with our estimate of analytical reproducibility (Supplementary Information). To estimate empirical reproducibility comparably for this, we fit a simple intercept-only model (removing the influence of age) with a logistic link function, to the probability that data is available for any study. We used the same estimate of analytical reproducibility as described above.

We conducted all analyses in R (v 3.6.1, R Core Team (2019)). We fit data decay and logistic regression models using the Stan engine, implemented in R with rstan (v 2.19.2, Stan Development Team (2019)). We used regularising priors to constrain parameters to empirically plausible values (McElreath, 2018). All results presented are summaries of 10,000 iterations of 4 chains for each model. We assessed convergence and mixing with the Gelman-Rubin diagnostic $\widehat{R}$, the estimate of the autocorrelation-adjusted number of samples (`n_eff`), and visual inspection of the trace plots of all chains. We provide details of all models described in the Supplementary Information, and data and analysis code for reproducing our reported results at https://github.com/rianaminocher/reproducibility-analysis.

2.2.1 Reproducibility protocols

Most papers present a mix of descriptive results ("we found materials online for 62 of 560 studies surveyed") and inferential results ("we estimated the half-life of human experimental data to be approximately 9 years"). Descriptive results are usually straightforward to reproduce, given data, while inferential results often involve some statistical modelling of the data, whether parametric, such as a logistic regression, or nonparametric, such as a Wilcoxon test, and thus require more analysis details.

When the data provided in the article or supplement was aggregated by subgroup or category to present results, we could not independently re-establish a reported descriptive result. We thus coded descriptive results as non-reproducible when we did not have access to individual, observation-level data, if they were simply stated or printed in a table in the article (analagous to failing to re-fit a model, while the model estimate is printed in text). This means that descriptive results usually fail to reproduce because of insufficient data, rather than unclear analysis details.

Importantly, our assessment of reproducibility did not involve evaluating the logic of a stated result; that is, whether we believed the evidence reported sufficiently supports the scientific claim. Instead, we followed the argument presented in the paper, assuming that the ability to create the same evidence reported ("analytical reproducibility") constitutes a successful reproduction. We did not critique the appropriateness of analysis choices for particular results, such as when we ourselves would have made a different choice about the selection or transformation of data, or methods applied. In each case, we followed the instructions exactly as presented. Thus, our assessment of reproducibility does not appraise the overall quality of research. Rather, we assert that such a reproduction procedure is necessarily prior to advocating an extension, alternative or a criticism of the quality of a published result or approach.

2.3 The reasonable researcher criterion

A major challenge to establishing a useful estimate of reproducibility is that reproduction attempts may fail for some researchers, while succeed for others. This may be because of a lack of access to necessary software (proprietary programs) or hardware (server clusters), or a lack of necessary skill or knowledge for a specialised analysis.

Given this, we consider when it is fair to label a result "non-reproducible". It follows that our expectations of access to software, hardware, skills and knowledge for any particular reproduction should be defined relative to the norms of a particular scholarly community.

Throughout our study, we attempted to remain conscious of the time, effort and skillset that a typical early-career researcher in the evolutionary behavioural sciences might invest in a reanalysis. We sampled papers published in English, we utilised institutional access to software and server clusters, we researched methods we were unfamiliar with, and we endeavoured to correspond with authors of publications in a formal, professional tone. This "reasonable researcher criterion" is central to our definition of a successful reproduction, at both the empirical or analytic reproduction stages. By our protocol, therefore, a result that can be recovered with a *more-than-reasonable* amount of effort is here considered the same as a result that cannot be reproduced at all.

3 Results

3.1 Empirical reproducibility

We were able to access materials online - in supplementary information, data repositories or in the article PDF - for 62 studies, approximately 11% of our total sample. We contacted authors of 473 of the remaining papers (84% of the total sample), and received a response of some kind in 315 cases (56%). Following our correspondence, we received materials for 105 publications (19%), bringing our full sample of papers with materials to reconstruct analyses to 167, that is, 30% of the full sample (Figure 1).

The availability of social learning data decays exponentially with time since publication (Figure 2). We estimated that the expected probability of finding material for any publication declines to half its current value (data half-life) within 5.86 years on average ($t(\frac{1}{2})$ 94.5% compatibility interval (CI): 4.88 - 6.85). Thus, the probability that data is available for studies published more 20 years ago is expected to be close to zero (Figure 2 (a), probability less than 0.01 at $t = 23$ years).

Estimating a type-specific data decay rate for human, non-human, experimental and observational data types, we found that the data decay rate of human experimental data is lower than expected for other data types (Figure 2). That is, the estimated expected half-life for human experimental data, of approximately 9 years (mean: 9.47, CI: 6.00 - 13.37), is substantially larger than half-lives ranging from 5 to 6 years for other data types (human observational: 5.34 (CI: 2.71 - 8.37), non-human observational: 4.92 (CI: 3.98 - 5.92), non-human experimental: 5.92 (CI: 3.80 - 8.06)).

Through email correspondence, we found that data were no longer available to share, either because of the data storage format or location, or because materials were not readily accessible to the author within the timeframe of our data request (anywhere between five weeks and five months). We received a response from the authors of 39 studies, indicating that they were willing to share the data from the

corresponding publication with us, but we did not receive the materials - we did not persist with data requests beyond the timeframe of the study. In 9 cases, authors indicated that they were unwilling to share research materials with us, implying that they may still have access to the relevant files. We have no information on the location or status of materials for 162 studies, a little under a third of our sample, having found no material online and receiving no reply from authors.

3.2 Analytical reproducibility

We attempted to reproduce the results of 40 randomly-selected studies, from the 167 that we categorised as empirically reproducible. We identified a total of 111 results within these 40 studies, with a median of 3 and a maximum of 6 results per study. We were able to reproduce 73 of 111 results, that is, produce at least a single line of quantitative evidence that confirmed the original stated finding. Accounting for unequal clustering of results within papers, we estimate the probability of reproducing a result from the social learning literature, conditional on materials being available, to be 79.6% (CI: 54.5 - 93.2%), (Figure 3).

Multiplying the posterior probability of obtaining materials for a study (29.9%, CI: 26.4 - 33.7%, Supplementary Information) with the posterior probability of reproducing the results of study (79.6%, CI: 54.5 - 93.2%), we estimate a combined probability of empirical and analytical reproducibility of social learning literature to be 23.8% (CI: 12.4 - 31.9%).

4 Discussion

Taken together, our results suggest that a reasonable effort to reproduce the published findings of social learning work will succeed for about one in every four attempts. That is, we expect materials to be inaccessible through reasonable request efforts for about 70 of 100 potential manuscripts; with the availability of data expected to decline exponentially with the age of the paper. Of 30 papers with accessible materials, we would expect a further 6 to fail to reproduce because of unclear data, ambiguous analysis methods or reanalyses disagreements. Thus, we expect a successful reproduction for 24 in 100 attempts, or just under one in four. From our analysis and our correspondence with authors in our sample, we attempt to characterise effective strategies for improving reproducibility of social learning research.

4.1 Data decay is the greatest challenge

The largest barrier to successful reproduction is the unavailability, incompleteness or ambiguity of datasets. We find a steep decline in data availability of about 11% per year, indicating a halving time of about six years. Essentially no data is available from the social learning literature before the year 2000. Furthermore, upon examining available data, we find the primary cause of analytical reproduction failure to be

incomplete and unusable datasets. Individual, community, and institutional solutions to extending data lifetime and usability remain major challenges for researchers.

Historically, sharing data in the behavioural sciences was neither normative nor feasible due to technological limitations. In the pre-Internet era, sharing data would have required creating and mailing copies on paper or physical storage media, in formats that are today mostly obsolete. Paper archives at research institutions were, for some respondents in our survey, no longer accessible and in a few cases had already been destroyed. Likewise, before the proliferation of open-source software tools, digital data was often stored in outdated or proprietary formats requiring expensive licenses for use. Frequently, replies to our requests included the phrase, "unfortunately, that study was X computers ago."

Currently, individual scientists employ a large suite of tools to extend the availability of data for future dissemination. The use of plaintext data formats ensures stable, long-term accessibility, and inexpensive or free hosting services are available through organisations like the Open Science Framework, GitHub, and Data Dryad. The FAIR guidelines (Findable, Accessible, Interoperable, and Reusable; (Wilkinson et al., 2016)) are a valuable starting place for cultivating robust practices for empirical projects, among others (Ihle et al., 2017; Culina et al., 2018; Whitlock, 2011). Indeed, today most researchers publishing in ecology and evolution are fluent in a wide array of computational tools (Fidler et al., 2017). Yet the range of skill and basic data science fluency between scientists could be further reformed; numerous resources are being developed, already available or are increasingly utilised to do so, such as Software Carpentry (`http://software-carpentry.org`), R for Data Science (Wickham and Grolemund, 2017), and The Practice of Reproducible Science (Kitzes et al., 2017).

Still, continuity of material support requires infrastructure and support beyond the roles of individual scholars. Many journals have adopted data sharing policies (Stodden et al., 2013; Simmons, 2017), and offer alternatives to online repositories such as the possibility to include data in supplementary materials (Hanson et al., 2011). However, these policies face three major challenges: proper enforcement (McCullough et al., 2008; Baker et al., 2014; Roche, 2017), including appraising files before publication (Hardwicke et al., 2018), expanded hosting for analysis code as well as data tables (Mislan et al., 2016), and encouraging standards of quality to authors and reviewers (Parker et al., 2018). If data is to be maintained as a shared community resource post retirement of individual scientists, research institutions maintain a responsibility to facilitate data lifetime, via access to storage platforms, in particular when datasets are large or numerous, or appointment of data management, to take requests when researchers retire.

4.2 Analytic clarity depends on data provenance

For 21 of 111 results we evaluated against available materials, we were unable to proceed with the reanalysis because we lacked a clear sense of what the available data represented. In some cases, the data was 'too raw'. Some authors could provide us with video or audio files, but not the descendent data tables coded or calculated from

these primary sources. While theoretically possible to re-process the source media, it was prohibitively difficult and failed our reasonable researcher criterion. In other instances, the data provided did not clearly correspond to the analyses presented in the article. The meaning of variables in datasheets can become impossible to decipher over time, given that we often change our description of variables for presentation in an article, or later create new, very similar variables. This is especially true for longitudinal datasets, when new data is continuously added, and could be conflated with older, different data. That data be archived along with a data dictionary or variable key in the language of publication (for our study, English) is not trivial practice.

Data is best curated for sharing as early as possible, ideally at the time of publication. Ostensibly, materials should be ready to share at the review stage of the publication process, as reviewers now regularly request access to the raw data as a part of publication review. When our requests were sent, some authors kindly invested time to prepare, document and tidy materials. This process can be prohibitive depending on the author's research schedule or career stage at the time of request. We therefore consider that if data were well-archived, well-documented and potentially made available with the paper publication, it should demand negligible time and effort to deal with a data request.

Beyond hindering reproducibility, poor data organisation greatly limits potential for alternative re-analyses. Data is easier to use for new purposes if it is organised in a 'long' format, i.e. a single entry per case, without aggregating information into categories or creating subsets of the data for analyses. If published data are to be re-used to ask new questions, maintaining *data provenance* is essential. That is, the entire history of the data and its origins, from data collection and transcription, to coding, transformation and analysis should be documented and stored. This can be facilitated by the use of a version control system (e.g. git) to archive data at various stages in its lifetime, to ensure we are able to return to different versions of the data if necessary.

4.3 More complex analyses are more difficult to reproduce

While data interpretability remains the primary cause of failure, we found the complexity of analyses themselves to be a critical barrier to analytic reproducibility. Understanding specific analysis decisions based on the descriptions in papers demanded a substantial amount of effort. Often we were able to devise a solution only by making inferential leaps. For six results, we were unable to reconstruct the analysis steps that led to the reported results, even with a clear understanding of the data, finally considering that further efforts would be unreasonable. For example, manuscripts sometimes referenced statistical techniques (often non-parametric tests) that we could not find through a search of standard reference literature, or described a quantitative method that was not consistent with the kind of result being reported. This tended to be especially true of papers that employed "point-and-click" programs like Excel and SPSS, which involve a number of consecutive calculations without record. In some of these cases, we were successfully able to use a different software or approach,

for example reproducing linear regressions in R that were originally performed with SPSS.

More recent analyses have been well-documented in statistical scripting languages such as R, Stata, Python, Mathematica and Matlab; we had access to analytical code for 11 of 40 studies. However, when studies involved complex or novel analytical techniques, a lack of analytical scripting was a major limitation to our reproductions. Several authors were hesitant to share their code with us, sometimes cautioning us that their scripts were likely too confusing or messy. That said, even when non-functional, code (in any programming language) greatly facilitated our ability to reconstruct results. Messy code is better than no code (Barnes, 2010), as it can document data manipulation or exclusion that is otherwise opaque, clarify the sequence and types of analyses conducted, and record any input or algorithm assumptions in the analysis. These details are likely too complex to compress comprehensively or reconstruct accurately from a written methods section, or even a detailed supplementary file. Moreover, an early commitment to share one's code may also affect the way that code is written, possibly encouraging the use of "linting" for code readability (Wickham, 2015), explanatory code comments, and other standard practices in open-source software development, thus facilitating reproducibility.

4.4 Attitudes towards reproducibility efforts are positive

Although only a minority of responses to our request included materials, we found a relatively high rate of response (68%), and strongly positive attitudes towards adopting reproducible practices. This is consistent with previous research on attitudes towards open science, that has indicated that scientists readily accept reproducible norms (Wallis et al., 2013) and reward reproducible practices (Kidwell et al., 2016). Younger researchers in particular are systematically more disposed to adopt open science practices (Campbell et al., 2019). Our finding that human experimental data has a longer lifetime than other data types in our sample may be indicative of research norms in the field of psychology; which may be more attuned to open science than other research areas, the "Replication Crisis" has also been called "the Credibility Revolution" (Vazire, 2018). The growing recognition of the importance of reproducibility is encouraging for researchers, particularly junior scholars, who are considering making such requests for materials from published work.

Some researchers, particularly of long-term observational datasets in our sample, expressed concerns to us about data re-use and subsequent misinterpretation upon publication of data. Data can be extremely costly to generate, leading researchers to justifiably feel that hard work of data collection should not be undervalued by making data freely available for re-use. Further, a fear that a wrong re-interpretation of published data can bolster false or contradictory analyses prevents some researchers from publicly sharing their data. However, if the goal of reproducibility is to permit an independent re-evaluation of published findings, we believe that data can and should be published with research articles, to a minimum community standard, potentially with restrictions on re-use, thus circumventing these issues for future work.

4.5 Implications for evolutionary anthropology & ecology

Our sample of studies cuts across interdisciplinary journals in which behavioural ecology and animal behaviour are commonly published, represents authors who work on diverse research questions, and samples a wide range of study species and disciplinary approaches. Thus, we have good reason to expect that the results from this survey generalise broadly to the field of behavioural ecology. Moreover, the decay rate of social learning data, while alarming, is not unique. Comparing our estimate of 30% empirical reproducibility, to similar findings of 38% (Vanpaemel et al., 2015), 32% (Hardwicke and Ioannidis, 2018), 26% (Wicherts et al., 2006), 19% (Vines et al., 2014); extending data lifetime appears consistently difficult to achieve, across disciplines. Our estimate of 80% analytical reproducibility is difficult to quantitatively compare against similar audits, which have found reproducibility of results to be anywhere between 83% (Andrew et al., 2015), 70% (Gilbert et al., 2012), and 1.1% (Stagge et al., 2019) of surveyed publications. This is likely because criteria for defining a successful reproduction effort, given materials, are currently ambiguous. Because the skillsets of a reasonable researcher can only be defined relative to the norms of a particular scholarly community, any reproducibility effort has to be defined in similar relative terms.

To assess analytical reproducibility, we focused on qualitative agreement with results stated in the publication abstract, findings that will likely be cited by future work. This is in contrast to reproducing specific estimates within a quantitative threshold, which may be feasible and more appropriate for a standardised type of analysis, (e.g. Andrew et al. (2015); Gilbert et al. (2012)). Furthermore, our characterisation of reproducibility in stages (Figure 1) is likely confounded, to the extent we cannot infer possible states of the data or result beyond the point of failure. We found the smallest number of analytical reproduction failures to be a disagreement between published results and reproduced findings. Thus, there is little evidence for substantial error or ambiguity in the analysis summaries. However, this low proportion needs to be qualified by the fact that all other failure states are logically prior, and would necessarily hide any potential cases of failure at this stage.

Correspondingly, to estimate a combined probability, we did not estimate a probability of analytical reproducibility of the set of papers which failed to reproduce empirically, that is, conditional on materials being unavailable. We assumed this to be zero. However, we could conversely assume that the probability an analysis for a paper without materials reproduces is high, if the probability of analytical reproducibility is uncorrelated with empirical reproducibility. Considering this, we attempted to incorporate further information about the overall sample that we extend our estimate of analytical reproducibility to, by assessing the influence of age of publication on probability of analytical reproducibility (Supplementary Information). That empirical reproducibility is heavily influenced by year of publication (Figure 2) is not unexpected, given the evolution of technology and norms over time. Yet we do not have reason to expect the same effect within the random subset of publications evaluated for analytical reproducibility. Accordingly, we found little to no effect of the year of publication on the probability of reproducing a result, albeit with low

confidence in this estimated effect (Supplementary Information). Thus, we do not use the influence of year to make a more informed estimate of combined probability across the sample.

We quantify the state of reproducibility within our field, based on a subset of published work, finding a combined probability of successful reproduction to be just under 25%. We caution that these estimates of reproducibility are a necessary summary of our findings, difficult to quantitatively compare across disciplines, and focus our discussion on facilitating reproducibility through data and code curation. We contribute evidence that data decays substantially over time, adding that the nature of the type of data archived may have influenced its survivability and shareability over time. Furthermore, we observe that data clarity remains the primary cause for analytical failure, even when materials are extant, further implying the need for improved data science skills, infrastructure and recognition of the importance of reproducibility.

4.6 Conclusion

Our survey of a literature within the field of evolutionary human science, the study of "social learning", shows that data preservation remains the primary barrier to research reproducibility. When materials are available for re-analysis, we estimate a high probability of recovering published results, with failure primarily due to incomplete or poorly curated data. Our correspondence with authors of social learning work indicated that attitudes towards reproducibility are generally positive, with researchers willing and committed to adopt reproducible research practices. Still, institutions and journals need to enforce standards for publication of datasets, while scientists continue to increase fluency in data carpentry skills. Recording data provenance carefully, particularly for observational, longitudinal datasets, as well as emphasising the need for additional documentation for complex, novel analyses, could extend the lifetime and re-use potential of data and methods. Reproducibility of empirical work is a minimum requirement for research in our field - a necessary precursor to replication, meta-analysis or further theory development. Taken together, our findings imply that the rapid adoption of reproducible norms and tools of data sharing will slow data decay tremendously, such that a study performed in a decade will show a very different decay rate, and correspondingly a much higher reproducibility rate, than estimated here.

4.7 Acknowledgements

Minocher designed the study, collected data, conducted analyses and wrote the paper. Beheim designed the study, conducted analyses and wrote the paper. McElreath provided crucial feedback on analyses and commented on the paper. Atmaca coordinated data collection and commented on the paper. Bavero collected data, corresponded with participating authors during the survey, and commented on the paper.

We thank Anne Büchner, Leonie Ette and Kristina Kunze for collecting data, and the Department of Human Behaviour, Ecology and Culture at the Max Planck

Institute for Evolutionary Anthropology for helpful discussion and feedback on the project at various stages.

References

Andrew, R. L., Albert, A. Y., Renaut, S., Rennison, D. J., Bock, D. G., and Vines, T. (2015). Assessing the reproducibility of discriminant function analyses. *PeerJ*, 3:e1137.

Baker, D., Lidster, K., Sottomayor, A., and Amor, S. (2014). Two years later: journals are not yet enforcing the arrive guidelines on reporting standards for pre-clinical animal studies. *PLoS Biology*, 12(1):e1001756.

Baker, M. (2016). Reproducibility crisis? *Nature*, 533(26):353–366.

Barnes, N. (2010). Publish your computer code: it is good enough. *Nature News*, 467(7317):753–753.

Begley, C. G. and Ioannidis, J. P. (2015). Reproducibility in science: improving the standard for basic and preclinical research. *Circulation Research*, 116(1):116–126.

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

Campbell, H. A., Micheli-Campbell, M. A., and Udyawer, V. (2019). Early career researchers embrace data sharing. *Trends in Ecology & Evolution*, 34(2):95–98.

Chang, A. and Li, P. (2015). Is economics research replicable? sixty published papers from thirteen journals say'usually not'. *Available at SSRN 2669564*.

Christensen, G. and Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3):920–980.

Culina, A., Baglioni, M., Crowther, T. W., Visser, M. E., Woutersen-Windhouwer, S., and Manghi, P. (2018). Navigating the unfolding open data landscape in ecology and evolution. *Nature Ecology & Evolution*, 2(3):420–426.

Devezer, B., Nardin, L. G., Baumgaertner, B., and Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLoS One*, 14(5).

Fidler, F., Chee, Y. E., Wintle, B. C., Burgman, M. A., McCarthy, M. A., and Gordon, A. (2017). Metaresearch for evaluating reproducibility in ecology and evolution. *BioScience*, 67(3):282–289.

Gilbert, K. J., Andrew, R. L., Bock, D. G., Franklin, M. T., Kane, N. C., Moore, J.-S., Moyers, B. T., Renaut, S., Rennison, D. J., Veen, T., et al. (2012). Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program structure. *Molecular Ecology*, 21(20):4925–4930.

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

Hanson, B., Sugden, A., and Alberts, B. (2011). Making data maximally available. *Science*, 331(6018):649.

Hardwicke, T. E. and Ioannidis, J. P. (2018). Populating the data ark: An attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. *PLoS One*, 13(8).

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.

Higginson, A. D. and Munafa, M. R. (2016). Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLoS Biology*, 14(11).

Huang, X., Hawkins, B. A., Lei, F., Miller, G. L., Favret, C., Zhang, R., and Qiao, G. (2012). Willing or unwilling to share primary biodiversity data: results and implications of an international survey. *Conservation Letters*, 5(5):399–406.

Ihle, M., Winney, I. S., Krystalli, A., and Croucher, M. (2017). Striving for transparent and credible research: Practical guidelines for behavioral ecologists. *Behavioral Ecology*, 28(2):348–354.

Ioannidis, J. P., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., Falchi, M., Furlanello, C., Game, L., Jurman, G., et al. (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics*, 41(2):149.

Ioannidis, J. P., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., Schulz, K. F., and Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet*, 383(9912):166–175.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., et al. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology*, 14(5).

Kitzes, J., Turek, D., and Deniz, F. (2017). *The practice of reproducible research: case studies and lessons from the data-intensive sciences*. University of California Press.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, 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.

Maxwell, S. E., Lau, M. Y., and Howard, G. S. (2015). Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, 70(6):487.

McCullough, B. D., McGeary, K. A., and Harrison, T. D. (2008). Do economics journal archives promote replicable research? *Canadian Journal of Economics/ Revue canadienne d'économique*, 41(4):1406–1420.

McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

McNutt, M. (2014). Journals unite for reproducibility.

Mislan, K., Heer, J. M., and White, E. P. (2016). Elevating the status of code in ecology. *Trends in Ecology & Evolution*, 31(1):4–7.

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., and Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1):0021.

Muthukrishna, M. and Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, 3(3):221–229.

Nissen, S. B., Magidson, T., Gross, K., and Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *eLife*, 5:e21451. arXiv: 1609.00494.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606.

OSF (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).

Parker, T. H., Griffith, S. C., Bronstein, J. L., Fidler, F., Foster, S., Fraser, H., Forstmeier, W., Gurevitch, J., Koricheva, J., Seppelt, R., et al. (2018). Empowering peer reviewers with a checklist to improve transparency. *Nature Ecology & Evolution*, 2(6):929.

Pashler, H. and Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6):528–530.

Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060):1226–1227.

Piwowar, H. A., Day, R. S., and Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS One*, 2(3):e308.

Plesser, H. E. (2018). Reproducibility vs. replicability: a brief history of a confused terminology. *Frontiers in Neuroinformatics*, 11:76.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roche, D. G. (2017). Evaluating Science's open-data policy. *Science*, 357(6352):654.

Schnitzer, S. A. and Carson, W. P. (2016). Would ecology fail the repeatability test? *BioScience*, 66(2):98–99.

Simmons, L. W. (2017). Guidelines for Transparency and Openness (TOP). *Behavioral Ecology*, 28(2):347–347.

Smaldino, P. E. and McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3(9):160384. arXiv: 1605.09511.

Stagge, J. H., Rosenberg, D. E., Abdallah, A. M., Akbar, H., Attallah, N. A., and James, R. (2019). Assessing data availability and research reproducibility in hydrology and water resources. *Scientific Data*, 6:190030.

Stan Development Team (2019). RStan: the R interface to Stan. R package version 2.19.2.

Stodden, V., Guo, P., and Ma, Z. (2013). Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals. *PLoS One*, 8(6):e67111.

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

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., and Storms, G. (2015). Are we wasting a good crisis? the availability of psychological research data after the storm. *Collabra*, 1(1):1–5.

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, 13(4):411–417.

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

Wallis, J. C., Rolando, E., and Borgman, C. L. (2013). If we share data, will anyone use them? data sharing and reuse in the long tail of science and technology. *PLoS One*, 8(7).

Whitlock, M. C. (2011). Data archiving in ecology and evolution: best practices. *Trends in Ecology & Evolution*, 26(2):61–65.

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

Wickham, H. (2015). *R packages: organize, test, document, and share your code*. O'Reilly Media, Inc.

Wickham, H. and Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, Inc., 1st edition.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3.

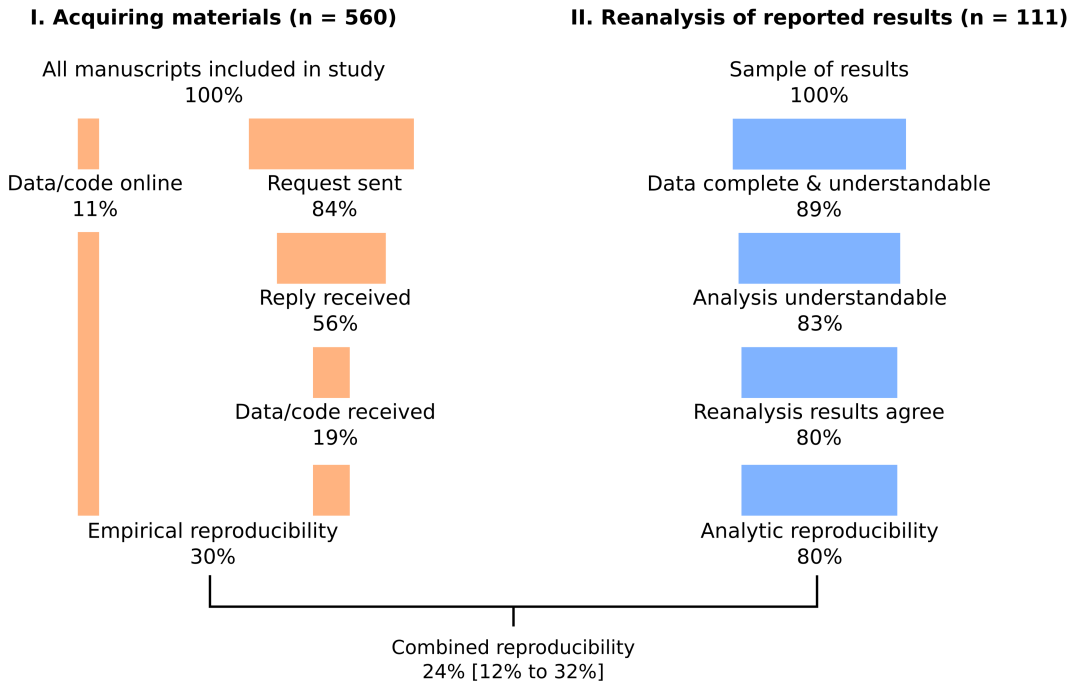

Figure 1: We evaluated empirical reproducibility, the availability of materials to reconstruct analyses, for 560 publications. We could access data online for 62 publications (11%). Contacting the authors of 473 of the remaining papers (84% of the total sample), we received a response of some kind in 315 cases (56%). Through this correspondence, we received materials for 105 publications (19%). Thus, we categorised a total of 167 studies with materials available, 30% of our initial sample. To evaluate analytical reproducibility, the ability to recreate published findings, given materials, we sampled 40 of these 167 publications randomly. From 111 results we identified, we estimated the probability of reproducing a publication given materials to be 80%. Causes of failure could be categorised sequentially, based on whether data is complete and understandable, analyses are understandable, and whether reproduced results differ from original results. The largest barrier to empirical reproducibility is the availability of data, rather than responsiveness of authors. Likewise, the largest barrier to analytical reproducibility is the usability or completeness of data, rather than opaque analyses or differing results.

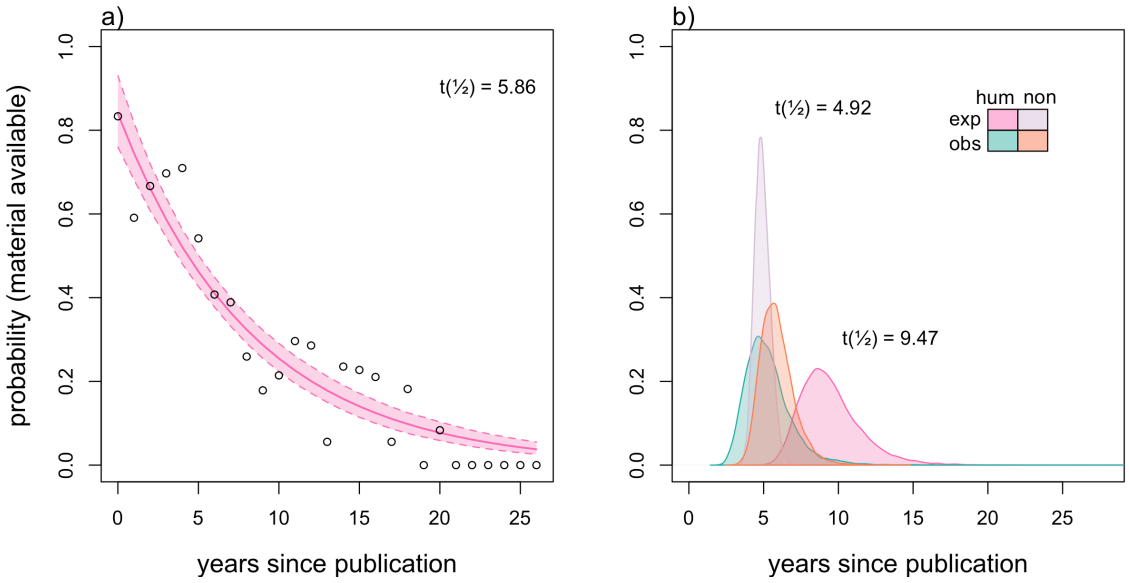

Figure 2: a) The predicted probability that material is available for any publication declines exponentially with time since publication, in years. The solid line plots the expected exponential decay curve. The shaded interval between dotted lines shows the 89% compatibility interval. The estimated decay rate corresponds to a half-life of 5.86 years, meaning that the probability at any time declines to half its present value within this period. The empty circles plot the raw proportion of studies, for each year, for which we obtained materials. b) The distribution of plausible half-life values, by type of study (human experimental; non-human experimental; human observational; non-human observational). Human experimental studies have a greater expected half-life than other study types.

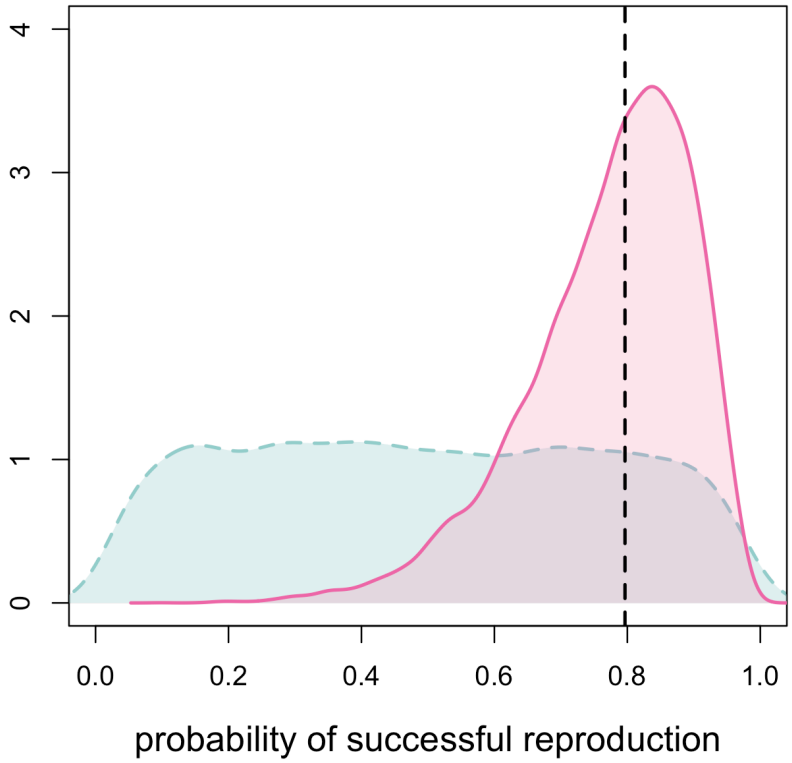

Figure 3: The posterior distribution of the probability of successfully reproducing a study (pink), conditional on materials being available, showing updating from the prior distribution (green). The dotted line plots the mean value of the intercept.