Extracting sunbeams from cucumbers: Computational reproducibility is essential for archaeological science

Ben Marwick1,✉

February 19, 2025

Abstract

Computational reproducibility is increasingly essential for advancing archaeological science, yet its implementation remains inconsistent across the discipline. This study presents a bibliometric assessment of archaeology’s status relative to other fields alongside a systematic review of computational reproducibility in published archaeological research. Analyzing nearly 10,000 articles, I identify trends in authorship, citation practices, and related metrics that position archaeology between the natural and social sciences. A review of reproducibility assessments for the Journal of Archaeological Science reveals persistent challenges, including missing data, unspecified dependencies, and inadequate documentation. To address these issues, I recommend to authors basic practical steps such as standardized project organization and explicit dependency documentation. Strengthening reproducibility will enhance archaeology’s scientific rigor and ensure the verifiability of research findings. This study underscores the urgent need for cultural and technical shifts to establish reproducibility as a cornerstone of rigorous, accountable, and impactful archaeological science.

1 Department of Anthropology, University of Washington

✉ Correspondence: [Ben Marwick <bmarwick@uw.edu>](mailto:bmarwick@uw.edu)

## Introduction

In their paper celebrating the 40th anniversary of this journal Torrence et al (2015) noted that reproducibility was an issue important to the reputation and sustainability of the discipline, and necessary for archaeological science to behave like a science. As part of the celebration of the 50th anniversary, and of Torrence’s leadership of the journal, my contribution revisits these topics of archaeology’s status as a science, this journal’s place in the landscape of archaeological science, and how the journal has responded to a growing recognition of the importance of reproducibility. I first present bibliometric evidence of the position of archaeology as a whole, and this journal in particular, in the sciences. Next I report on the journal’s progress in supporting reproducible research, and my work doing a new kind of peer review for JAS, one that evaluates the computational reproducibilty of the research submitted for publication. Finally, I analyse twelve months of reproducibility reviews to identify common weaknesses in the ways archaeologists are working currently, and provide simple recommendations for researchers to overcome these and contribute to the improvement computational reproducibility in archaeological science.

The question of archaeology’s status as a science usually comes up in the context of what the discipline should or should not be. One of the first landmarks in tackling this question is the debate published by *Antiquity* between classical archaeologist Jacquetta Hawkes and palaeoanthropologist Glynn Isaac. Hawkes (1968), advocating a humanistic archaeology, was concerned that scientific approaches to archaeology were causing researchers to be “swamped by a vast accumulation of insignificant, disparate facts, like a terrible tide of mud, quite beyond the capacity of any man to contain and mould into historical form”. More optimistic about the integration of science and archaeology, Isaac (1971) counters that “New levels of precision in presenting data and in interpreting them can surely lead to briefer and more interesting technical reports as well as providing the basis for more lively literary portrayals of what happened in prehistory. Expanding on Isaac’s perspective, Binford (1962) argued that archaeology should operate as a science after the model proposed by philosopher Carl Hempel, which prescribed hypothesis-driven approaches, leading to generalizable laws of human behavior. Counter-arguments came from numerous directions, notably Hodder (1985) who rejected the quest for generalisations and instead argued that archaeology should be subjective and reflective, focussed on symbolic and relational meanings of material culture and the historical particularity of past human cultures. These debates, and the many more similar ones summarised by Martinón-Torres and Killick (2013), have become a genre in archaeological writing that can be characterized as mostly based on personal observations, microscopic dissections of a handful of cherry-picked case studies of good or bad practice, and discussion of various philosophers and sociologists.

What has been missing from these debates is a macroscopic observation of what the majority of archaeologists are actually doing, and an empirical comparison to a broad spectrum of relatively harder and softer disciplines. At the ‘hard’ end of the spectrum (e.g. physics and chemistry), scholars more typically share a large set of established set of theories, facts, and methods, facilitating fairly rapid agreement on the validity and significance of new results. At the ‘soft’ end of the spectrum (e.g. economics and psychology), the set of theories, facts, and methods on which there is widespread consensus is smaller, and agreement is slower and less frequently reached about the significance of new findings and the continuing relevance of previous work. Hardness and softness is a controversial distinction, in part because it is often used to imply a rank order of disciplines that encodes legitimacy, productivity, perceived value to society, and worthiness of funding (Cole, 1983; Editors, 2012). However, independent of these value judgments, empirical analysis of scholarly articles does indicate a spectrum of variation in practice linked to differing degrees of consensus in a discipline, for example in approaches to data visualisation (Cleveland, 1984; Smith et al., 2000). Similarly, quantitative analysis of the frequency of positive results (ie. full or partial support for a research hypothesis) in publications is significantly correlated with hardness, consistent with a model where researchers in harder fields more readily accept any result their research produces, while those in softer fields have more freedom to choose which theories and hypotheses to test and how to interpret results (Fanelli, 2010). The hard-soft spectrum is also evident in surveys of how researchers view their own work relative to those in other fields (Biglan, 1973).

## How to measure the hardness or softness of a science?

To objectively quantify the relative hardness or softness of archaeology, as an evaluation of its status as a science, and the place of this journal in context of other archaeology journals, I take a bibliometric approach. This approach is based on Fanelli and Glänzel (2013), who examined the hardness and softness of 12 disciplines using scholarly publication parameters. Fanelli and Glänzel (2013) found a spectrum of statistically significant variation in bibliometric variables from the physical to the social sciences, with papers tending at the softer end of the spectrum tending to have fewer co-authors, use less substantive titles, have longer texts, cite older literature, and have a higher diversity of sources. In Fanelli and Glänzel (2013)’s analysis harder sciences include Space Science, Physics, Chemistry, less hard sciences include social sciences include Psychiatry, Psychology, Economics, Business and General Social Sciences, and the Humanities define the soft end of the spectrum. Following Fanelli and Glänzel (2013), I quantify the number of authors, length of article, relative title length, age of references, and diversity of references for a large sample of peer-reviewed journal articles.

These parameters are useful because of how they signify consensus in a research community. A larger number of authors on a paper reflects collaboration of people working together on a common goal. Collaborators have specialized roles, each of whom has the ability to study a part of the problem with high accuracy and detail, with harder fields having larger groups of collaborators (Zuckerman and Merton, 1972). Reflecting this collaboration group size, harder disciplines tend to have higher average numbers of authors on papers. Article length has an inverse correlation with field hardness. In low-consensus, or softer, fields, papers must be longer to present justification, nuance and contextualization of results. The number of substantive and informative words in an article’s title tends to be positively correlated with article length in harder disciplines (Yitzhaki, 2002; Yitzhaki, 1997), reflecting a focus on empiricism and efficiency that is characteristic of high-consensus disciplines. While Yitzhaki (2002) removed stop-words (e.g. prepositions, articles, conjunctions, etc.) to calculate article length, in order to generate results for comparison with Fanelli and Glänzel (2013) I follow his method of dividing the total word count of the article title by the total number of pages of the article to compute relative title length.

The age of works cited has long been used as a measure of a field’s hardness (Börner, 2010; Moed et al., 1998), based on the assumption that harder fields assimilate new results more rapidly that softer fields (Price, 1970). I calculated a recency of references index for each article (also known as the Price index), which is the proportion of all cited works that were published in the five years preceding the paper. The diversity of references

While Fanelli and Glänzel (2013) analysed papers published in a single year (2012), I found only 303 papers for that same year, and 70% of papers in the sample published after that date. To make efficient use of the available data and ensure robust representation from different areas of archaeology, including those with lower frequencies of journal article publication, I analysed 9697 papers published during 1975-2025. This sample was collected from Clarivate’s Web of Science database by first selecting the Web of Science category ‘Archaeology’ and the Document type ‘article’ (n = 28,871). To focus on journals of broad relevance to most archaeologists, and that are representative of substantial communities of practice, I then filtered the results to keep only articles published in the top-ranking 25 journals according to their h-indices as reported by Clarivate’s Journal Citation Indicator. Finally I excluded journals with less than 100 articles in the database, resulting in 20 journals.

The entire R code (R Core Team, 2024) used for all the analysis and visualizations contained in this paper is included in the Supplementary Online Materials at https://doi.org/xxx/xxx/xxx to enable re-use of materials and improve reproducibility and transparency (Marwick, 2017). All the figures, tables, and statistical test results presented here can be independently reproduced with the code and data in this compendium (Marwick et al., 2018). The R code is released under the MIT license, the data as CC-0, and figures as CC-BY, to enable maximum re-use.

## Results

### How does archaeology compare to other fields?

|  |
| --- |
| Figure 1: Distributions of article characteristics hypothesised to reflect the level of consensus. The boxplot shows the distribution of values of archaeology articles. The thick black line in the middle of the boxplot is the median value, the box represents the inter-quartile range (the range between the 25th and 75th percentiles, where 50% of the data are located), and individual points represent outliers. The thin grey horizontal lines in each boxplot indicate the median values computed by Fanelli and Glanzel (2013), where p = physics, s = social sciences, h = humanities. ln denotes the natural logarithm, or logarithm to the base e. |

[Figure 1](#fig-compare-other-fields) shows the distribution of bibliometric variables for archaeology in context of data from other fields presented by Fanelli and Glänzel (2013). The most striking indicator of archaeology as a hard science is the number of authors, where it is between the social sciences and physics. Archaeology is a close fit with the social sciences in relative title length. It is between the the social sciences and humanities in recency of references and diversity of references. The clearest indicator of archaeology as a soft science is article length where it is similar to the humanities. Overall archaeology does not sit squarely at either end of the hard-soft spectrum. It is generally not a harder science than the social sciences, with the exception of collaborator group sizes.

### How has the hardness of archaeology varied over time?

|  |
| --- |
| Figure 2: Distribution of article characteristics for archaeology articles over time. Data points represent individual articles. The colour of the points indicates if the overall trend is toward softer (orange) or harder (green) Generalized Additive Models were computed to fit the lines summarising the relationships between the variables and the time series. |

[Figure 2](#fig-change-over-time) shows how the bibliometric indicators of field hardness have changed of time for archaeology articles. By two measures, the number of authors and relative title length, archaeology has become increasingly harder over time. On the other hand, three metrics indicate that archaeology has become softer (diversity of references, article length and recently of references). Although all the relationships are statistically significant, generally these temporal trends are very weak with low slope values, indicating very slow change over time. Similarly the r-squared values are very low, demonstrating that much of the variability in these metrics is independent of time.

The most striking change over time is in the increase in the number of authors, which has the highest r-squared value of these metrics. One interesting detail evident in [Figure 2](#fig-change-over-time) is the increase in the range of diversity of references after about 2010. This may be due to some broader changes in academic publishing around this time, such as moves to digital-first continuous publishing, new journals appearing (e.g. *Archaeological and Anthropological Sciences* in 2009 and *Journal of Island & Coastal Archaeology* in 2010), and non-archaeology journals becoming more relevant to archaeologists. For example, *PLOS ONE* received its first impact factor in 2010 and in 2011 Nature’s *Scientific Reports* began publishing (Malashichev, 2017).

### How do archaeology journals vary in hardness?

|  |
| --- |
| Figure 3: Panels A-E: Variation in bibliometric indicators of hardness for 20 archaeological journals. The journals are ordered for each indicator so that within each plot, the harder journals are at the top of the plot and the softer journals are at the base. Panel F shows a bar plot that is the single consensus ranking computed from all five variables, using the Borda Count ranking algorithm. Panel G is a biplot of the first and second principal components of a PCA computed on the journal means of the five bibliometric variables. |

[Figure 3](#fig-variation-by-journal) shows the distribution of our bibliometric variables of hardness for each of the 20 journals in the sample. Overall agreement between these bibliometric variables in ranking these journals on a hard-soft spectrum is moderate to strong, with a Kendall’s coefficient of concordance (Wt) value of 0.64 (in a 0-1 range, where 1 is perfect agreement) and a p-value of 2.67 x 10-06. Panel F of [Figure 3](#fig-variation-by-journal) shows an overall consensus ranking of all journals in the sample. In this consensus ranking the *Journal of Archaeological Science* among the top five archaeology journals for hardness. It is placed at the harder end of the hard-soft spectrum especially by the number of pages and relative title length, and to lesser degrees by the number of authors and recency of references. However, according to the diversity of references, the *Journal of Archaeological Science* is at the middle of the spectrum.

The *Journal of Cultural Heritage* is the only journal that consistently ranks as hard across all variables, occurring in the top five journals for all five metrics. This journal primarily publishes materials science and computational analyses related to conservation and preservation of historic objects in museums and other collections. Authors of papers in recent issues have affiliations with museums, cultural heritage programs, and chemistry, engineering, and physics departments at European and Chinese universities. Notably, papers in this journal typically do not engage in questions or debates about past human behaviour or culture. The absence of these questions in research published in this journal makes it an outlier here, since these questions are central to a common definition of archaeology as ‘cultural anthropology of the past’, a phrase first found in Leroi-Gourhan (1946) and repeated in widely-used contemporary undergraduate textbooks such as Renfrew et al. (2024). Most archaeologists would likely be surprised at the decision by Clarivate to include the *Journal of Cultural Heritage* in their category of archaeology journals, leading to this result in [Figure 3](#fig-variation-by-journal) where the hardest archaeology journal publishes papers that are not very archaeological at all because they do not engage with anthropological topics.

The *Journal of Archaeological Research* is notable because it consistently ranks as soft, ranking as the softest journal for four of our five bibliometric variables. This is a predicable result for a review journal, which is a distinct type of journal dedicated to summarizing, analyzing, and synthesizing existing research in a particular field. The stated aim of the *Journal of Archaeological Research* is to ‘bring together the most recent international research summaries on a broad range of topics and geographical areas’ (Feinman and Parkinson, 2024). A typical article is a long single-authored synthesis of archaeology in a region or on a topic. As the only review journal in this sample, this is a stark contrast to the other journals here that present original research findings, and like the *Journal of Cultural Heritage*, may be considered an outlier in this sample.

The PCA results in panel G show of [Figure 3](#fig-variation-by-journal) that PC1 captures most of the variance in the metrics (71%) and is reasonable proxy for the hard-soft spectrum, with *Journal of Cultural Heritage* representing the hard extreme on the right and *Journal of Archaeological Research* representing the soft extreme on the left. The distribution of PC1 values is skewed left, with most of the journals concentrated at the harder end of the spectrum. The distribution of PC2 values reveals additional structure to the data and can be roughly separated into generalist journals in the negative range (e.g. *American Antiquity*, *Antiquity*, *Advances in Archaeological Practice*), and more specialised journals in the positive range (e.g. *Environmental Archaeology*, *Geoarchaeology*, *Archaeologial Research in Asia*, *Journal of Island and Coastal Archaeology*). The *Journal of Archaeological Science* sits about midway between these two groups, reflecting its relevance to specialised communities of practice in archaeology, as well as general

## Reproducibility: A key measure of how scientific a field is

This macroscopic perspective derived from an analysis of the ways thousands of archaeologists communicate their research has produced a complex picture of archaeology as a science. In the context of a broad spectrum of other research areas, archaeologists behave like social scientists. We are harder than typical social scientists in tending to form larger groups of collaborators more often, and softer in sometimes writing longer articles than more resemble humanities scholarship. The outlook for the future of archaeology is also complex, with three out of five of the bibliometric variables trending towards more humanistic styles of working, but the discipline showing more extreme values in some metrics towards both hard and soft sciences after about 2010. Among archaeology journals, we see distinct communities of practice reflected in the PCA results that are very close together on the hard-soft spectrum, but have minor differences in their communication styles, perhaps due to differences in writing traditions inherited from parent disciplines such as geology and biology.

While these bibliometric variables provide several interesting insights into the status of archaeology as a science, via measurement of consensus, and are important for moving the debate beyond discussions of a small number of case studies, they miss a crucial factor that separates scientific practice from non-science. This is reproducibility, which, according to a report by the US National Science Foundation (Cacioppo et al., 2015), “refers to the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results… Reproducibility is a minimum necessary condition for a finding to be believable and informative.” The importance of this factor can be traced to Irish chemist Robert Boyle (1627-1691), best known for his experiments with vacuum pumps (Shapin and Schaffer, 2011). Boyle was concerned about the secrecy common among experimentalists in the 17th century and aimed to shift the culture from valuing direct in-person witnessing of scientific demonstrations towards meticulous written communications that were detailed enough to enable a reader to successfully undertake the experiment themselves, independent of the original author.

With many disciplines making increasing use of computationally intensive analyses in recent years there has been renewed interest in reproducibility (LeVeque et al., 2012). In part, this is because computationally intensive research is difficult to communicate within the constraints of the methods section of a traditional journal article — the reader needs also needs the computer code written by the original authors, not just the article text. There is also the broader context of rising pressure to publish in prestigious journals and intense competition for funds that create strong incentives for malpractice in research (Edwards and Roy, 2017). These two factors have led to widespread concerns of a reproducibility crisis in many fields (Baker, 2016). Estimates of scientific reproducibility in several fields confirm the extent of this problem. Empirical replications of 100 studies published in three psychology journals found that 36% of replications had statistically significant results, compared to 96% of the original studies (Open Science Collaboration, 2015). Similar empirical replications of large numbers of social science studies and experiemental economics studies successfully replicated 61% and 62% of their target studies respectively (Camerer et al., 2018; Camerer et al., 2016).

Similarly bleak results come from measurements specifically of the reproducibility of computational analyses of scientific studies. An attempt at reproducing the computational results of 204 papers in *Science* resulted in success in reproducing the findings for 26% (Stodden et al., 2018). The computational results of two out of 41 geoscience papers could be fully reproduced on the first attempt (Konkol et al., 2019). In the biomedical field, code in 1,203 out of 27,271 (4%) notebooks associated with 3,467 publications could be run without errors (Samuel and Mietchen, 2024). Statisticians could reproduce 15% of 93 papers (Xiong and Cribben, 2023). Economists have been especially active in researching computational reproducibility, with studies indicating successful reproduction of results using code and data provided by authors for 30% of 67 papers (Chang and Li, 2015), 14% of 203 papers (Gertler et al., 2018), 44% of 152 papers (Herbert et al., 2021), 30% of 419 articles (Fišar et al., 2024), and 28% of 168 papers (Pérignon et al., 2024). These efforts confirm that the reproducibility of published research is widely recognised as a cornerstone of rigorous science, and work on evaluating how successful a research community is at generating reproducible results has become a distinctive and important meta-research activity in many fields.

How does archaeology compare to these other fields in terms of reproducibility? Empirical reproducibility has long been valued in field archaeology. Throughout the history of archaeology, well-known sites have been repeatedly revisited to test old hypotheses with new evidence or methods, for example, Olduvai Gorge (Tanzania), Cahokia (USA), Çatalhöyük (Turkey), and Madjedbebe (Australia). Similarly among many experimental archaeologists, empirical reproducibility is a key concern, for example in lithic use-wear identification (Hayes et al., 2017) and the measurement of lithics (Pargeter et al., 2023). To what extent does this concern extend to computational reproducibility among archaeologists?

## How reproducible is archaeology? Investigating computational reproducibility

In 2024 the *Journal of Archaeological Science* introduced a new kind of peer review that has provided an opportunity to tackle this question about computational reproducibility in archaeology. In January 2024 I accepted the position of ‘Associate Editor for Reproducibility’ (AER) for JAS and conducted reproducibility reviews of submissions that mentioned programming languages such as R or Python in the methods sections. A reproducibility review examines the code and data used to generate the results presented in the paper, and attempts to run the authors’ code to reproduce the their results (see Editors (n.d.) for more details about this process). This new AER role is based on similar positions (i.e. ‘data editor’ or ‘reproducibility editor’) that journals in economics (Vilhuber, 2019), statistics (Wrobel et al., 2024), astronomy (Muench, 2023), ecology (Bolnick et al., 2022), and environmental studies (Rosenberg et al., 2021) have had, in some cases for over a decade. In 2024, three archaeology journals, in addition to JAS, added AERs to their editorial communities: *Advances in Archaeological Practice* (Marwick, 2024, one paper reviewed), *Journal of Field Archaeology* (Farahani, 2024, two papers reviewed), and *American Antiquity* (Martin, 2024, no papers reviewed).

|  |
| --- |
| Figure 4: Summary of reproducibility reviews for JAS. A: Primary software used for the computational analysis reported in a manuscript. B: Computational or statistical method used by the authors (GMM = geometric morphometrics, Frequentist = hypothesis tests such as chi-square and ANOVA, AI/ML = artifical intellegence and machine learning, including neural networks and deep learning, Network = statitical analysis of social networks, 3D = analysis of 3D data such as artefact models, Composition = compositional analysis of artefacts). C: Locations where authors deposited their code and data files. D: Issues that prevented the reproducible review from succeeding on the first attempt. E: Relationship between software used and issues that make research irreproducible. |

At the time of writing (January 2025) we have completed 47 reproducibility reviews of 25 manuscripts submitted to JAS. Of these, 11 have been published in JAS to date. Seven of these eleven papers fully passed the reproducibility review, resulting in a success rate, by one measure, of 63%. Four of the seven papers could be fully reproduced on my first attempt, the others required additional input from the authors. For comparison with reproducibility studies in other fields reported above, the seven fully reproducible papers should be divided by the 25 reviewed for reproducibility, resulting in a 28% success rate. Expanding the denominator to include the total number of research articles published in JAS from May 2024 (when the first article to pass the reproducibility review, Herskind and Riede (2024), was published) to January 2025 (n = 97), which are articles that could have been eligible for reproducibility review had the authors used an open source programming language (e.g. instead of commercial software such as Microsoft Excel or SPSS, etc.) the success rate is 7%. In any case, the computational reproducibility of archaeological research is generally on the low end of the distribution of values available from a variety of hard and soft sciences.

[Figure 4](#fig-aer-summary) shows a summary of basic characteristics of the 25 articles that have been through the reproducibility review process so far. The most commonly used software is R, followed by Python. Results generated with proprietary or closed-source software are out of scope for reproducibility reviews. Several distinct types of analyses are well-represented in this sample, especially geometric morphometry, network statistics, and analyses using artificial intelligence or machine learning algorithms (this includes deep learning and neural networks). Most authors are sharing their code and data files via Zenodo, a non-profit generic research data repository hosted by CERN that accepts any file format and freely assigns all publicly available uploads a DOI to make the files easily and uniquely citable (Peters et al., 2017). In this same category of DOI-issuing, research-grade repositories is OSF (the Open Science Foundation), Figshare, and university repositories. GitHub, a commercial service owned by Microsoft is a code hosting platform that is convenient for collaboration, is also popular among JAS authors, but is a problematic choice because does not offer DOIs or the same commitments to long-term availability as Zenodo. Some authors attached their code and data as journal article supplementary files, but this is a poor choice for long-term availability because these files are typically renamed and converted to different formats during the article production process, making it difficult or impossible for a reader to combine the code and data to reproduce the results.

Panels D and E of [Figure 4](#fig-aer-summary) summarise the common issues that resulted in irreproducible results. The most common issue was an incomplete compendium. This ranges from missing data files down to missing lines of code. In most cases this can attributed to accidental carelessness, with the exception of two cases where data was unavailable due to licensing restrictions. Unspecified or under-specified dependencies is another common issue that prevents code from running. This refers to the software packages in addition to R or Python that an author used to do specialized analyses and visualisations, (e.g. dplyr for R or numpy for Python). If an author does not clearly specify the name and version number of the packages that they used for their analysis, it can be very time-consuming or impossible to correctly identify these because many packages have functions with similar names, and functions in any one package can change they way they behave as the developers update their package. Other reasons why papers failed the reproducibility review is that the paths to data files were incorrectly specified (likely a result of the author reorganising their compendium after completing their analysis, or omitting data files from the materials submitted for review), and errors returned by functions, which have diverse causes.

## How to improve the computational reproducibility of archaeology?

Despite the relatively small number of reproducibility reviews reported on here, there are patterns of common issues that point to a small set of simple tasks authors can do that have high potential to increase reproducibility. The problem of incomplete materials can be tackled in several basic ways. First, authors should use a simple and logical folder structure to organise their code and data to be as self-contained as possible (ie. not requiring files outside of the project folder). There are many excellent, simple, and widely-used templates that authors can choose from, e.g. Marwick et al. (2018), Figueiredo et al. (2022), Greenfeld and Community (2023), Cooper and Hsing (2017) and Wilson et al. (2017)

Second, for analyses that are not highly time-consuming (which was over 90% of the sample), authors should re-run their code more than once, and ideally not on the same computer (e.g. by another co-author of the paper), before submission to confirm everything works as expected. This would ensure portability and likely also prevent issues relating to path and function errors. Complex and time-consuming analyses should use pipeline or workflow management tools, e.g. GNU Make, Luigi, Snakemake, or Targets, to document the relationship of the files and folders in a machine-readable format and simplify running and re-running code by others (Landau, 2021; Wratten et al., 2021).

Third, authors should include in their compendium a README document that describes to readers the folders and files contained in the project. The README file is typically the first file that a reader will look at in a compendium so it should include brief instructions to guide the user to a successful reproduction of the original results (e.g. what order to run the code files in). A README should also briefly describe the contents of the compendium, where other necessary files can be obtained (e.g. data files not included in the compendium), and the key software packages needed and the version numbers that the authors used.

Fourth, authors should ensure clear, direct and obvious connections between their code and the results they present in their paper. One simple way to do this is to have one code file for each figure and table, and name the code files with the figure or table number and some key words in the caption. Another way some authors are accomplishing this is by using literate programming tools, such as Quarto and Jupyter notebooks (Allaire and Dervieux, 2024; Kluyver et al., 2016) that enable the research narrative and code for data analysis to be woven together in one document. Quarto was a popular tool among the JAS papers in the reproducibility review sample, for example, Vernon and Ortman (2024) and Ragno (2024) wrote their entire manuscripts using Quarto. These four simple steps have high potential to substantially increase the computational reproducibility of archaeological science.

Our finding that dependencies are a common cause of irreproducible results is consistent with previous studies that have identified this as a widespread weakness in communicating computationally intensive research (Samuel and Mietchen, 2024; Trisovic et al., 2022). In our sample, issues relating to dependencies are strongly associated with the use of Python. One possible reason for this is that relative to R, Python uses more package managers, more environments, and deeper dependency chains with more complex inter-dependencies that change more rapidly (Decan et al., 2019; Decan et al., 2016; Korkmaz et al., 2020). Another reason may be that there is a bigger and more established community of R users in archaeology (Batist and Roe, 2024; Schmidt and Marwick, 2020) that highly values code that is easy for others to reuse and has evolved practices to effectively communicate dependencies (e.g. Bilotti et al., 2024; Will and Rathmann, 2025).

The simplest way for archaeologists, to improve here is to write the names and version numbers of the software and packages they used in their README file, as we see in Herskind and Riede (2024) and Monna et al. (2024). For more complex research projects, i.e. those using five or more packages or machine learning algorithms, authors should use dependency management tools to keep track of the packages and version numbers needed to reproduce their results. This is an active area of development, and while there are many tools currently available, the most robust and widely used include renv for R Ragno (2024) and conda and poetry for Python (Anaconda, 2023; Crasta et al., 2023).

A more comprehensive solution, and the leading best practice for managing dependencies in many computationally intensive fields using Python in particular, is to include a Dockerfile in the compendium (Moreau et al., 2023). This is a set of a machine- and human-readable instructions that enables a user to recreate the author’s computational environment (including those requirements beyond the R or Python packages) on another computer (Nüst et al., 2020). Dockerfiles are gradually being adopted by archaeologists, see Crema et al. (2024) and Liao et al. (2024) for examples, and most of our reproducibility reviews include a recommendation that the author include a Dockerfile to manage complex dependencies efficiently.

## Conclusion

In the classic satirical novel *Gulliver’s Travels* (1726) by Irish writer Jonathon Swift Gulliver visits the fictional Grand Academy of Lagado in Balnibarbi, a caricature of the Royal Society of London, and meets several researchers working on wildly impractical projects, including one attempting to extracting sunbeams out of cucumbers. This is usually interpreted as a subversive anti-colonial parody depicting institutionalized research as an absurd fund-raising activity with no practical benefits to society (Alff, 2014; Nicolson and Mohler, 1937). It has also been used as a metaphor for the difficulty of getting insights from data tables in scholarly publications (Feinberg and Wainer, 2011). This metaphor is relevant in our current age of computationally intensive research, where extracting useful code and data from the publication of a typical computational study can be as frustrating and fruitless as extracting sunbeams from a cucumber.

I have presented a bibliometric analysis on the status of archaeology as a science, showing distinct disunity that is increasing over time. On average we generally behave as social scientists, with some elements in common with harder sciences. These observations are consistent with (**Lakatos1978model?**)’s of a research program as a central foundation of irrefutable core assumptions complemented by a set of hypotheses, models, and methods that are adjusted, modified, or replaced by day-to-day research. Archaeology consists of multiple programs like this, with distinct and sometimes non-overlapping sets of core assumptions, such as processual and post-processual archaeology. Some programs are more amenable to reproducibility, while others offer insights through alternative methods. Among the programs that depend on quantitative methods to assess hypotheses and models, if they are to continue to progress through increased consensus through the accumulation of reliable facts and methods, it is essential for researchers to take computational reproducibility seriously. Computers have become a central field and laboratory instrument for much of our work, so we have an ethical duty to document how we change our data as it flows through silicon just as carefully as we document the operating parameters of a mass spectrometer or any other field or laboratory instrument. However, the current state of quantitative archaeology, with most researchers not using open source code, is comparable to the secrecy of alchemy prior to the emergence of chemistry. Abandoning this habit of secrecy in favour of transparency and reproducibility is vital if we are to avoid a future where our journals are filled with pretty pictures that the reader has no hope of repeating or adapting in their own work. Computational reproducibility must be considered a minimum requirement for evaluating the integrity and usefulness of empirical results.

Computational reproducibility is not a panacea, results that are fully reproducible can contain errors and fraud. It is no guarantee of code quality, or that statistics have been used appropriately (cf. Vaiglova 2025; Crema 2025), or that data management is consistent with FAIR and CARE principles (Carroll et al., 2021). It is also time-consuming for authors to ensure their computational work can be reproduced, and for reviewers to evaluate. In an professional environment where job security and career progression is often associated with pressure to publish many high-impact papers, demands for authors to spend time on reproducibility, resulting in less time for publishing more papers, may seem frustrating. This may seem especially unfair to early career researchers on short-term contracts, who may feel the goalposts for career success are being moved and that they are being asked to do more work their graduate training has not prepared them for. This highlights the need for a culture shift among senior archaeological scientists to value reproducibility in hiring and promotion decision-making. Professors must contribute to this shift by nurturing a culture of reanalysis and reproducibility in their teaching, for example by using replication assignments and by training students in the most current best practices and tools for reproducible research, such as R and Python (Marwick et al., 2020).

# References

Alff, D., 2014. Swift’s solar gourds and the rhetoric of projection. Eighteenth-Century Studies 47, 245–260.

Allaire, J., Dervieux, C., 2024. [Quarto: R interface to ’quarto’ markdown publishing system](https://cran.r-project.org/web/packages/quarto/index.html).

Anaconda, Inc., 2023. [Conda: A cross-platform package and environment manager](https://docs.conda.io/).

Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. Nature 533, 452–454. <https://doi.org/10.1038/533452a>

Batist, Z., Roe, J., 2024. Open Archaeology, Open Source? Collaborative practices in an emerging community of archaeological software engineers. Internet Archaeology. <https://doi.org/10.11141/ia.67.13>

Biglan, A., 1973. The characteristics of subject matter in different academic areas. Journal of Applied Psychology 57, 195–203. <https://doi.org/10.1037/h0034701>

Bilotti, G., Kempf, M., Oksanen, E., Scholtus, L., Nakoinz, O., 2024. Point Pattern Analysis (PPA) as a tool for reproducible archaeological site distribution analyses and location processes in early iron age south-west Germany. PLOS ONE 19, e0297931. <https://doi.org/10.1371/journal.pone.0297931>

Binford, L.R., 1962. Archaeology as anthropology. American antiquity 28, 217–225.

Bolnick, D., Vines, T., Montgomerie, B., 2022. [Ensuring data and code archive quality: Why and how?](https://comments.amnat.org/2022/03/ensuring-data-and-code-archive-quality.html) (Editorial). From the Editor’s Desk of The American Naturalist.

Börner, K., 2010. [Atlas of science](https://mitpress.mit.edu/9780262014458/atlas-of-science/). The MIT Press.

Cacioppo, J.T., Kaplan, R.M., Krosnick, J.A., Olds, J.L., Dean, H., 2015. Social, behavioral, and economic sciences perspectives on robust and reliable science. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences 1.

Camerer, C.F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., Wu, H., 2016. Evaluating replicability of laboratory experiments in economics. Science 351, 1433–1436. <https://doi.org/10.1126/science.aaf0918>

Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B.A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., Wu, H., 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour 2, 637–644. <https://doi.org/10.1038/s41562-018-0399-z>

Carroll, S.R., Herczog, E., Hudson, M., Russell, K., Stall, S., 2021. Operationalizing the CARE and FAIR Principles for Indigenous data futures. Scientific Data 8, 108. <https://doi.org/10.1038/s41597-021-00892-0>

Chang, A.C., Li, P., 2015. Is economics research replicable? Sixty published papers from thirteen journals say “usually not.” Finance and Economics Discussion Series. <https://doi.org/10.17016/FEDS.2015.083>

Cleveland, W.S., 1984. Graphs in scientific publications. The American Statistician 38, 261–269.

Cole, S., 1983. The hierarchy of the sciences? American Journal of Sociology 89, 111–139. <https://doi.org/10.1086/227835>

Cooper, N., Hsing, P.-Y., 2017. [A Guide to Reproducible Code in Ecology and Evolution](https://nhm.openrepository.com/handle/10141/622618). British Ecological Society.

Crasta, S., Gannstyf, D., Developers, P., 2023. [Poetry: Python dependency management and packaging](https://python-poetry.org/).

Crema, E.R., Bloxam, A., Stevens, C.J., Vander Linden, M., 2024. Modelling diffusion of innovation curves using radiocarbon data. Journal of Archaeological Science 165, 105962. <https://doi.org/10.1016/j.jas.2024.105962>

Decan, A., Mens, T., Claes, M., 2016. On the topology of package dependency networks: A comparison of three programming language ecosystems, in: Proccedings of the 10th European Conference on Software Architecture Workshops. pp. 1–4.

Decan, A., Mens, T., Grosjean, P., 2019. An empirical comparison of dependency network evolution in seven software packaging ecosystems. Empirical Software Engineering 24, 381–416. <https://doi.org/10.1007/s10664-017-9589-y>

Editors, n.d. [Reproducibility at journal of archaeological science - journal of archaeological science | ScienceDirect.com by elsevier](https://www.sciencedirect.com/journal/journal-of-archaeological-science/about/reproducibility-at-journal-of-archaeological-science/reproducibility-at-jas).

Editors, 2012. A different agenda. Nature 487, 271–271. <https://doi.org/10.1038/487271a>

Edwards, M.A., Roy, S., 2017. Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition. Environmental engineering science 34, 51–61.

Fanelli, D., 2010. “Positive” Results Increase Down the Hierarchy of the Sciences. PLOS ONE 5, e10068. <https://doi.org/10.1371/journal.pone.0010068>

Fanelli, D., Glänzel, W., 2013. Bibliometric Evidence for a Hierarchy of the Sciences. PLOS ONE 8, e66938. <https://doi.org/10.1371/journal.pone.0066938>

Farahani, A., 2024. Reproducibility and archaeological practice in the journal of field archaeology. Journal of Field Archaeology 49, 391–394. <https://doi.org/10.1080/00934690.2024.2391623>

Feinberg, R.A., Wainer, H., 2011. Extracting sunbeams from cucumbers. Journal of Computational and Graphical Statistics 20, 793–810. <https://doi.org/10.1198/jcgs.2011.204a>

Feinman, G., Parkinson, W., 2024. [Aims and Scope: Journal of Archaeological Research](https://link.springer.com/journal/10814). SpringerLink.

Figueiredo, L., Scherer, C., Cabral, J.S., 2022. A simple kit to use computational notebooks for more openness, reproducibility, and productivity in research. PLOS Computational Biology 18, e1010356. <https://doi.org/10.1371/journal.pcbi.1010356>

Fišar, M., Greiner, B., Huber, C., Katok, E., Ozkes, A.I., Collaboration, M.S.R., 2024. Reproducibility in management science. Management Science 70, 1343–1356.

Gertler, P., Galiani, S., Romero, M., 2018. How to make replication the norm. Nature 554, 417–419.

Greenfeld, A.R., Community, C., 2023. Cookiecutter: A command-line utility that creates projects from project templates.

Hawkes, J., 1968. The Proper Study of Mankind. Antiquity 42, 255–262. <https://doi.org/10.1017/S0003598X00034451>

Hayes, E.H., Cnuts, D., Lepers, C., Rots, V., 2017. Learning from blind tests: Determining the function of experimental grinding stones through use-wear and residue analysis. Journal of Archaeological Science: Reports 11, 245–260. <https://doi.org/10.1016/j.jasrep.2016.12.001>

Herbert, S., Kingi, H., Stanchi, F., Vilhuber, L., 2021. [The reproducibility of economics research: A case study](https://www.banque-france.fr/en/publications-and-statistics/publications/reproducibility-economics-research-case-study) (Working Paper No. 853). Banque de France.

Herskind, L.L.P., Riede, F., 2024. A computational linguistic methodology for assessing semiotic structure in prehistoric art and the meaning of southern scandinavian mesolithic ornamentation. Journal of Archaeological Science 165, 105969. <https://doi.org/10.1016/j.jas.2024.105969>

Hodder, I., 1985. Postprocessual archaeology. Advances in archaeological method and theory 1–26.

Isaac, G.L., 1971. Whither archaeology? Antiquity 45, 123–129. <https://doi.org/10.1017/S0003598X00069283>

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., others, 2016. Jupyter notebooks–a publishing format for reproducible computational workflows, in: Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS press, pp. 87–90.

Konkol, M., Kray, C., Pfeiffer, M., 2019. Computational reproducibility in geoscientific papers: Insights from a series of studies with geoscientists and a reproduction study. International Journal of Geographical Information Science 33, 408–429. <https://doi.org/10.1080/13658816.2018.1508687>

Korkmaz, G., Kelling, C., Robbins, C., Keller, S., 2020. Modeling the impact of Python and R packages using dependency and contributor networks. Social Network Analysis and Mining 10, 1–12. <https://doi.org/10.1007/s13278-019-0619-1>

Landau, W.M., 2021. The targets r package: A dynamic make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. Journal of Open Source Software 6, 2959.

Leroi-Gourhan, A., 1946. Archéologie du pacifique-nord: Matériaux pour l’étude des relations entre les peuples riverains d’asie et d’amérique, Travaux et mémoires de l’institut d’ethnologie. Institut d’ethnologie, Paris.

LeVeque, R.J., Mitchell, I.M., Stodden, V., 2012. Reproducible research for scientific computing: Tools and strategies for changing the culture. Computing in Science and Engineering 14, 13–17. <https://doi.org/10.1109/mcse.2012.38>

Liao, L., Sun, Z., Liu, S., Ma, S., Chen, K., Liu, Y., Wang, Y., Song, W., 2024. Applying a mask r-CNN machine learning algorithm for segmenting electron microscope images of ceramic bronze-casting moulds. Journal of Archaeological Science 170, 106049. <https://doi.org/10.1016/j.jas.2024.106049>

Malashichev, Y., 2017. From open access to open science. Biological Communications 3–5.

Martin, D.L., 2024. Editor’s corner. American Antiquity 89, 163–164. <https://doi.org/10.1017/aaq.2024.31>

Martinón-Torres, M., Killick, D., 2013. [Archaeological Theories and Archaeological Sciences](http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199567942.001.0001/oxfordhb-9780199567942-e-004).

Marwick, B., 2024. Introducing the associate editor of reproducibility. Advances in Archaeological Practice 12, 61–62. <https://doi.org/10.1017/aap.2024.15>

Marwick, B., 2017. Computational reproducibility in archaeological research: Basic principles and a case study of their implementation. Journal of Archaeological Method and Theory 24, 424450. <https://doi.org/10.1007/s10816-015-9272-9>

Marwick, B., Boettiger, C., Mullen, L., 2018. Packaging data analytical work reproducibly using r (and friends). The American Statistician 72, 80–88. <https://doi.org/10.1080/00031305.2017.1375986>

Marwick, B., Wang, L.-Y., Robinson, R., Loiselle, H., 2020. How to Use Replication Assignments for Teaching Integrity in Empirical Archaeology. Advances in Archaeological Practice 8, 78–86. <https://doi.org/10.1017/aap.2019.38>

Moed, H.F., Van Leeuwen, T.N., Reedijk, J., 1998. A new classification system to describe the ageing of scientific journals and their impact factors. Journal of Documentation 54, 387–419.

Monna, F., Navarro, N., Esin, Y., Rolland, T., Wilczek, J., Dumont, L., Magail, J., Allard, A.-C., Chateau-Smith, C., Mongush, C., Byrynnay, S., Alibert, P., 2024. Studying seriality in material culture by geometric morphometricsgold wild boars from the arzhan-2 barrow, tuva. Journal of Archaeological Science 169, 106021. <https://doi.org/10.1016/j.jas.2024.106021>

Moreau, D., Wiebels, K., Boettiger, C., 2023. Containers for computational reproducibility. Nature Reviews Methods Primers 3, 1–16. <https://doi.org/10.1038/s43586-023-00236-9>

Muench, A., 2023. The roles of data editors in astronomy. Science Editor 46, 8–10. <https://doi.org/10.36591/SE-D-4601-04>

Nicolson, M., Mohler, N.M., 1937. The scientific background of swift’s voyage to laputa. Annals of Science 2, 299–334.

Nüst, D., Sochat, V., Marwick, B., Eglen, S.J., Head, T., Hirst, T., Evans, B.D., 2020. Ten simple rules for writing Dockerfiles for reproducible data science. PLOS Computational Biology 16, e1008316. <https://doi.org/10.1371/journal.pcbi.1008316>

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349, aac4716. <https://doi.org/10.1126/science.aac4716>

Pargeter, J., Brooks, A., Douze, K., Eren, M., Groucutt, H.S., McNeil, J., Mackay, A., Ranhorn, K., Scerri, E., Shaw, M., Tryon, C., Will, M., Leplongeon, A., 2023. Replicability in Lithic Analysis. American Antiquity 88, 163–186. <https://doi.org/10.1017/aaq.2023.4>

Pérignon, C., Akmansoy, O., Hurlin, C., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Menkveld, A.J., Razen, M., Weitzel, U., 2024. Computational reproducibility in finance: Evidence from 1,000 tests. The Review of Financial Studies 37, 3558–3593. <https://doi.org/10.1093/rfs/hhae029>

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., Gorraiz, J.I., 2017. Zenodo in the spotlight of traditional and new metrics. Frontiers in Research Metrics and Analytics 2. <https://doi.org/10.3389/frma.2017.00013>

Price, D., 1970. Citation measures of hard science, technology and nonscience. Communication among scientists and engineers 3–22.

R Core Team, 2024. R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.R-project.org/., Vienna, Austria.

Ragno, R., 2024. Sheep and goats taxonomic abundance trends in 1st millennium CE southern italy: Multilevel bayesian modelling of NISP datasets. Journal of Archaeological Science 171, 106068. <https://doi.org/10.1016/j.jas.2024.106068>

Renfrew, C., Bahn, P., DeMarrais, E., 2024. Archaeology: Theories, methods, and practice, 9th ed. Thames & Hudson, New York.

Rosenberg, D.E., Jones, A.S., Filion, Y., Teasley, R., Sandoval-Solis, S., Stagge, J.H., Abdallah, A., Castronova, A., Ostfeld, A., Watkins, D., 2021. Reproducible Results Policy. Journal of Water Resources Planning and Management 147, 01620001. <https://doi.org/10.1061/(ASCE)WR.1943-5452.0001368>

Samuel, S., Mietchen, D., 2024. Computational reproducibility of jupyter notebooks from biomedical publications. GigaScience 13, giad113. <https://doi.org/10.1093/gigascience/giad113>

Schmidt, S.C., Marwick, B., 2020. Tool-driven revolutions in archaeological science 3, 1832. <https://doi.org/10.5334/jcaa.29>

Shapin, S., Schaffer, S., 2011. Leviathan and the air-pump: Hobbes, boyle, and the experimental life. Princeton University Press.

Smith, L.D., Best, L.A., Stubbs, D.A., Johnston, J., Archibald, A.B., 2000. Scientific graphs and the hierarchy of the sciences:: A latourian survey of inscription practices. Social Studies of Science 30, 73–94. <https://doi.org/10.1177/030631200030001003>

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, 2584–2589. <https://doi.org/10.1073/pnas.1708290115>

Torrence, R., Martinón-Torres, M., Rehren, Th., 2015. Forty years and still growing: *Journal of Archaeological Science* looks to the future. Journal of Archaeological Science, Scoping the future of archaeological science: Papers in honour of richard klein 56, 1–8. <https://doi.org/10.1016/j.jas.2015.03.001>

Trisovic, A., Lau, M.K., Pasquier, T., Crosas, M., 2022. A large-scale study on research code quality and execution. Scientific Data 9, 60. <https://doi.org/10.1038/s41597-022-01143-6>

Ushey, K., Wickham, H., 2025. [Renv: Project environments](https://rstudio.github.io/renv/).

Vernon, K.B., Ortman, S.G., 2024. A method for defining dispersed community territories. Journal of Archaeological Science 170, 106048. <https://doi.org/10.1016/j.jas.2024.106048>

Vilhuber, L., 2019. Report by the AEA data editor. AEA Papers and Proceedings 109, 718–729. <https://doi.org/10.1257/pandp.109.718>

Will, M., Rathmann, H., 2025. Exploring the utility of unretouched lithic flakes as markers of cultural change. Scientific Reports 15, 1571. <https://doi.org/10.1038/s41598-025-85399-z>

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., Teal, T.K., 2017. Good enough practices in scientific computing. PLOS Computational Biology 13, e1005510. <https://doi.org/10.1371/journal.pcbi.1005510>

Wratten, L., Wilm, A., Göke, J., 2021. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nature Methods 18, 1161–1168. <https://doi.org/10.1038/s41592-021-01254-9>

Wrobel, J., Hector, E.C., Crawford, L., McGowan, L.D., Silva, N. da, Goldsmith, J., Hicks, S., Kane, M., Lee, Y., Mayrink, V., others, 2024. Partnering with authors to enhance reproducibility at JASA. Journal of the American Statistical Association 1–3.

Xiong, X., Cribben, I., 2023. The state of play of reproducibility in statistics: An empirical analysis. The American Statistician 77, 115–126. <https://doi.org/10.1080/00031305.2022.2131625>

Yitzhaki, M., 2002. Relation of the title length of a journal article to the length of the article. Scientometrics 54, 435–447. <https://doi.org/10.1023/A:1016038617639>

Yitzhaki, M., 1997. Variation in informativity of titles of research papers in selected humanities journals: A comparative study. Scientometrics 38, 219–229. <https://doi.org/10.1007/BF02457410>

Zuckerman, H., Merton, R.K., 1972. Age, aging, and age structure in science. Higher Education 4, 1–4.