Journal of Archaeological Science 180 (2025) 106281

Contents lists available at ScienceDirect

Journal of Archaeological Science

journal homepage: www.elsevier.com/locate/jas

Is archaeology a science? Insights and imperatives from 10,000 articles and a year of reproducibility reviews

Ben Marwick

Department of Anthropology, University of Washington, United States

A B S T R A C T

The status of archaeology as a science has been debated for decades and influences how we practice and teach archaeology. This study presents a novel bibliometric assessment of archaeology's status relative to other fields using a hard/soft framework. It also presents a systematic review of computational reproducibility in published archaeological research. Reproducibility is a factor in the hardness/softness of a field because of its importance in establishing consensus. 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 survey of reproducibility reviews 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. 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 reproducibility 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 of 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. Drawing on a different group of philosophers, Smith (2017) argues for archaeology more specifically as a social science. Bevan (2015) proposes that floods of digital data are reconfiguring our analytical agendas and support empirical and inductive inference. Counter-arguments to archaeology as a science come 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

This article is part of a special issue entitled: <Special issue full title (fetch from PTS); guest edited by editor's name (fetch from PTS)> published in <journal name>. Example: This article is part of a special issue entitled: Next Generation Archaeological Science: Papers in Honour of Robin Torrence; guest edited by Efthymia Nikita and Marcos Martinón-Torres.

E-mail address: bmarwick@uw.edu.

https://doi.org/10.1016/j.jas.2025.106281
Received 19 February 2025; Received in revised form 28 May 2025; Accepted 29 May 2025
Available online 18 June 2025

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 (Biglan, 1973). 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. In sum, the hard-soft status is defined by the amount of consensus in a field, and the speed at which consensus is reached on new knowledge (Fanelli and Glänzel, 2013). Hardness and softness is a controversial distinction, in part because it is sometimes used to imply a rank order of disciplines that encodes legitimacy, productivity, perceived value to society, and worthiness of funding (Cole, 1983; Editors, 2012). Another criticism is that it may be more of an emergent product of social and institutional processes rather than intrinsic differences in method or consensus (Latour, 1987). On the other hand, analyzing the characteristics that lead to the hard-soft distinction can be useful for understanding the diversity of academic inquiry, such as how different fields approach knowledge and differences in what counts as evidence and modes of argument, where fruitful collaborations might be possible due to shared methods and assumptions, and for curriculum design to structure courses appropriately based on a field's typical ways of knowing (Becher and Trowler, 2001).

Independent of these value judgments, empirical analysis of scholarly articles does support the hard-soft concept as 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).

2. How to measure the hardness or softness of a science?

To objectively quantify the diversity of modern archaeological practice across a scale of relative hardness or softness, 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 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's (2013) analysis harder sciences include Space Science, Physics, Chemistry, softer sciences include social sciences (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 specialised 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. While article length is constrained by journal requirements, leaving individual authors with little freedom to vary, journal requirements are typically set by editors who are professional archaeologists keen to tailor their journal to be attractive and relevant to other members of the discipline. Thus journal requirements for article length will reflect the norms of the discipline at any given time. The number of substantive and informative words in an article's title tends to be positively correlated with article length in harder disciplines (Yitzhaki, 1997, 2002), 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 their 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 is a similar indicator, with papers in harder fields having a higher concentration of more specific citations because more knowledge is taken for granted as core knowledge that does not need citing (Skilton, 2006). Conversely, softer fields have less knowledge taken for granted, a smaller core of facts that do not need citing, and thus a higher diversity of citations.

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 at https://doi.org/10.5281/zenodo.14897252 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.

3. Results

3.1. How does archaeology compare to other fields?

Fig. 1 shows the distribution of bibliometric variables for archaeology in the 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 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.

3.2. How has the hardness of archaeology varied over time?

Fig. 2 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 Fig. 2 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). The appearance of Google Scholar in 2004, increasing the discoverability of many works for many researchers, may have also contributed to this increase in diversity of references.

3.3. How do archaeology journals vary in hardness?

Fig. 3 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 2.67×10^{-06}. Panel F of Fig. 3 shows an overall consensus ranking of all journals in the sample. In this consensus

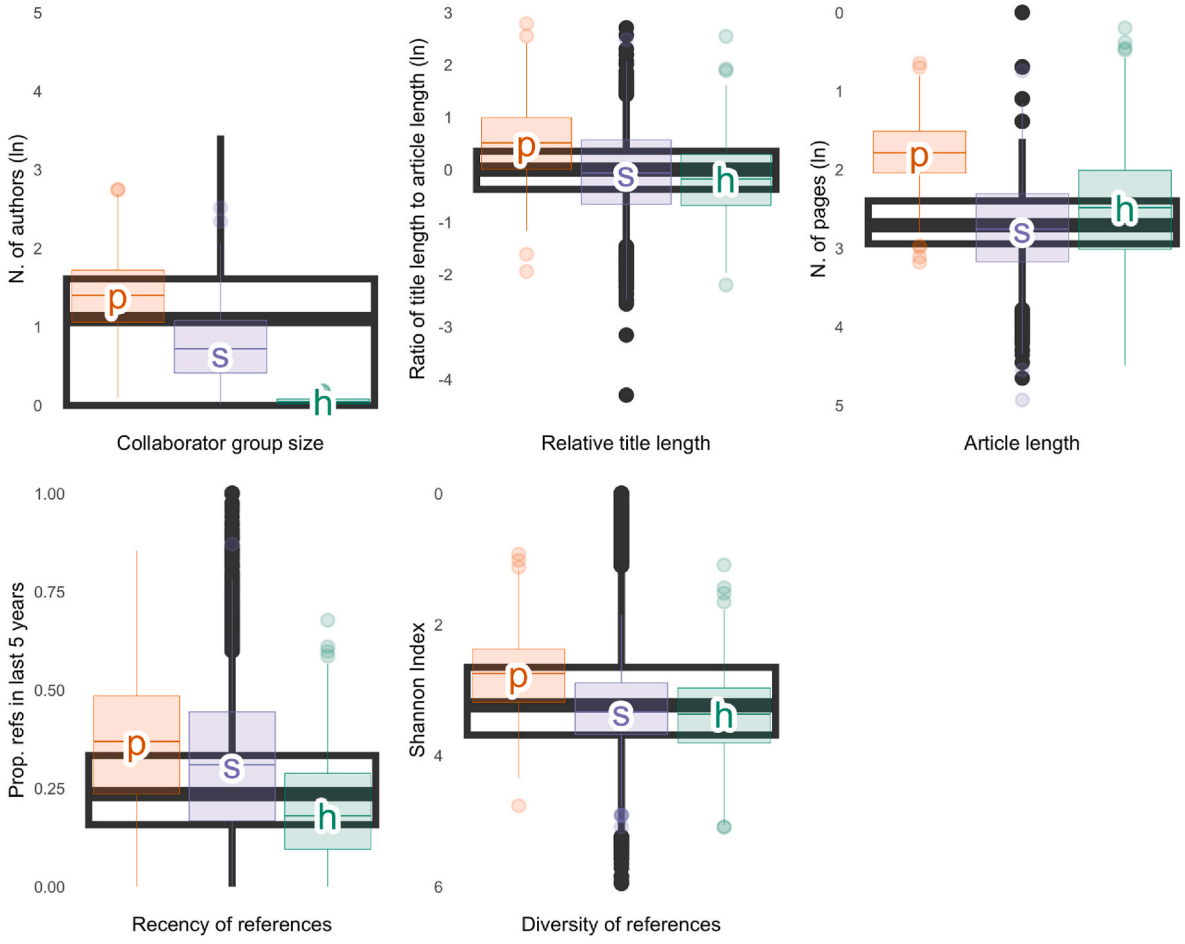

Fig. 1. Distributions of article characteristics hypothesised to reflect the level of consensus. The boxplot shows the distribution of values of archaeology articles. The thick black line in the middle of the boxplot is the median value, the box represents the inter-quartile range (the range between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

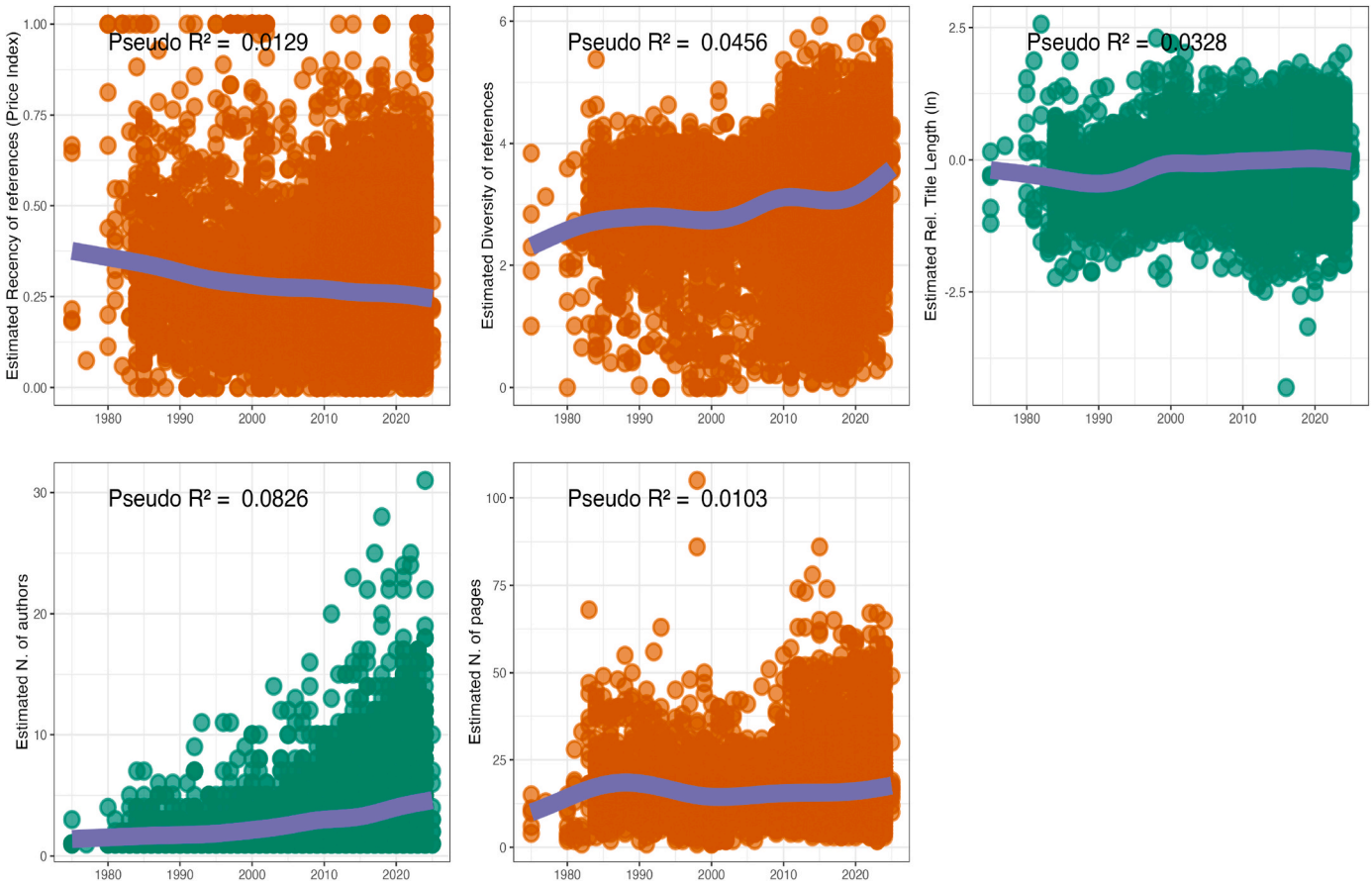

Fig. 2. Distribution of article characteristics for archaeology articles over time. Data points represent individual articles. The colour of the points indicates if the overall trend is toward softer (orange) or harder (green). Bayesian Generalized Additive Models were computed to fit the lines summarising the relationships between the variables and the time series. For recency of references a Zero-One Inflated Beta distribution family was used with a logit link function. For diversity of references and relative title length (ln) the Gaussian family was used with the canonical identity link function. For the number of authors pages the Negative Binomial family was used with a standard log link function. Further details about the model specifications and diagnostics are available at https://doi.org/10.5281/zenodo.14897252. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

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 behavior 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 Fig. 3 where the hardest archaeology journal publishes papers that are not very archaeological because they do not engage with anthropological topics.

The *Journal of Archaeological Research* is notable for consistently ranking as soft; it was 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 summarising, 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 Fig. 4 show that PC1 captures most of the variance in the metrics (71 %) and is a 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 variables that contribute to variation in PC1 are title length, number of pages, and the diversity of references. Journals with higher PC1 values have articles with longer titles, fewer pages, and less diverse reference lists. The distribution of PC1 values is skewed left, with most of the journals concentrated at the harder end of the spectrum. Variation in PC2 is influenced by the number of authors and recency of references. The distribution of PC2 values reveals additional structure to the data and can be roughly separated into generalist journals in the negative range of the PC2 axis (e.g. *American Antiquity*, *Antiquity*, *Advances in Archaeological Practice*), characterised by fewer authors and more recent references. In the positive range of the PC2 axis

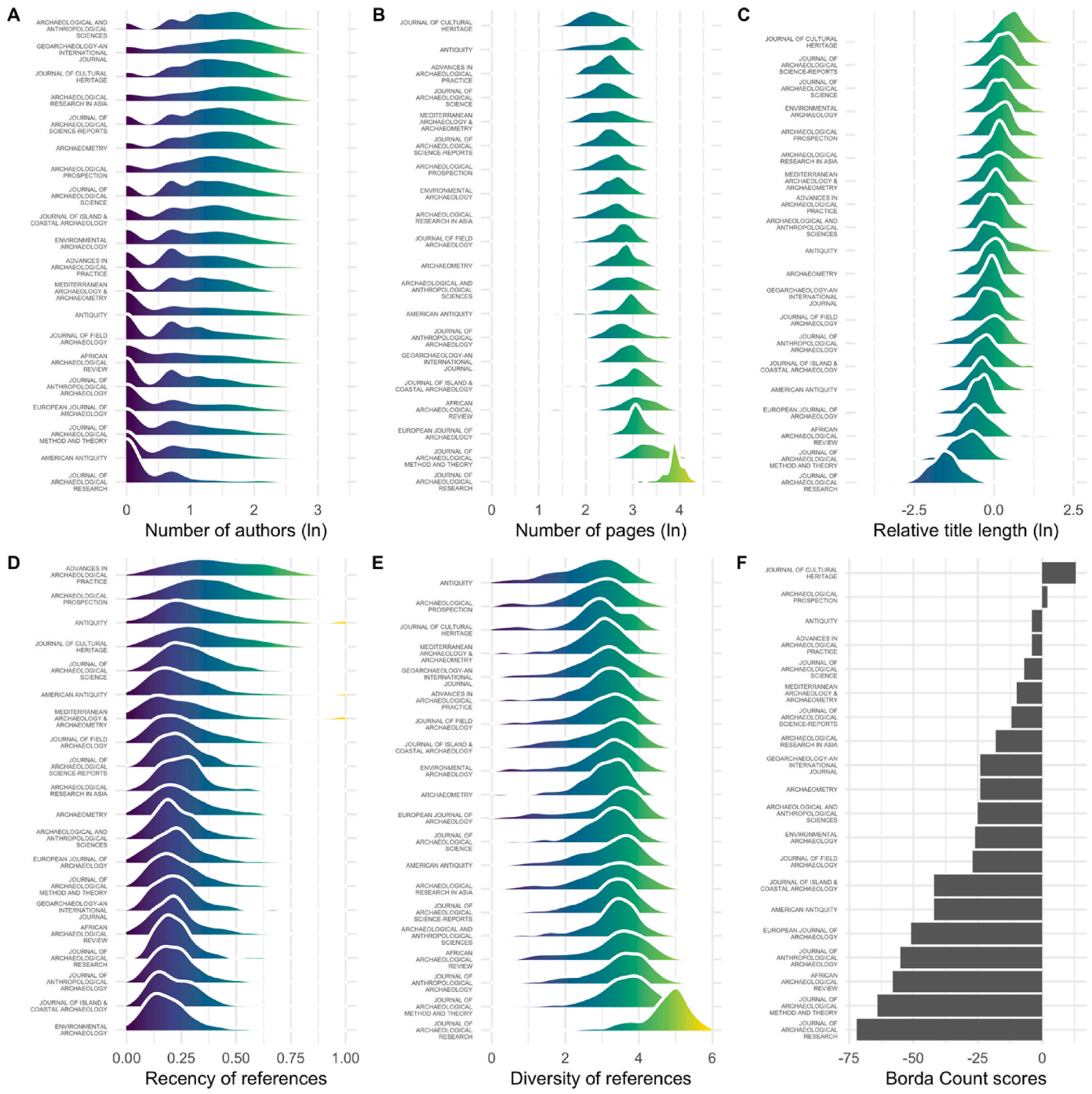

Fig. 3. Panels A–E: Variation in bibliometric indicators of hardness for 20 archaeological journals. The journals are ordered for each indicator so that within each plot, the harder journals are at the top of the plot and the softer journals are at the base. Panel F shows a bar plot that is the single consensus ranking computed from all five variables, using the Borda Count ranking algorithm.

are more specialised journals, characterised by higher numbers of authors and less recent references cited (e.g. *Environmental Archaeology*, *Geoarchaeology*, *Archaeological 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 both specialised and generalist communities of practice in archaeology.

4. 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 that 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

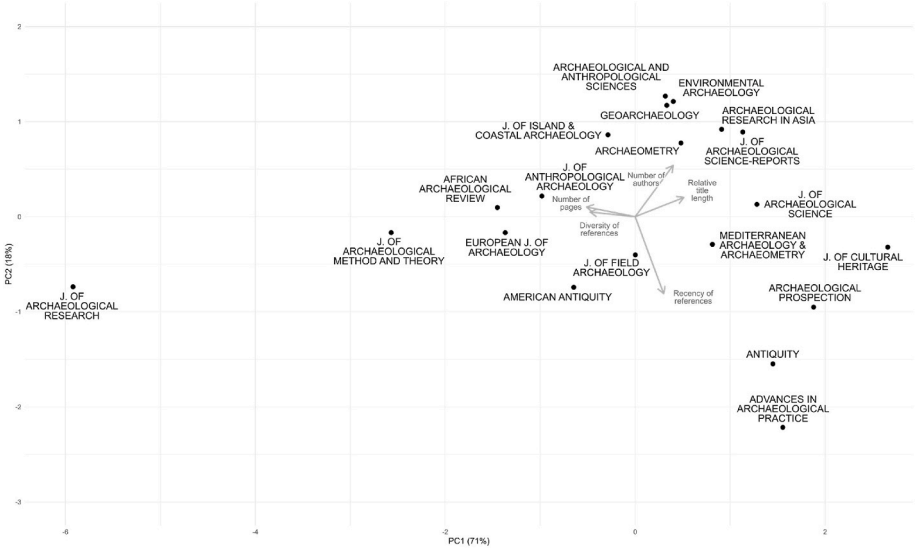

Fig. 4. Biplot of the first and second principal components of a PCA computed on the means of the five bibliometric variables for each journal in the sample. The arrows represent the correlation between each original variable and the principal components. The direction and length of the arrows indicate how strongly each variable contributes to each component.

spectrum, but have minor differences in their communication styles, perhaps due to cultural differences in writing traditions inherited from parent disciplines such as geology and biology (Becher and Trowler, 2001).

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 for 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." Scientific reproducibility is a factor contributing to the hardness of a field. Specifically, reproducibility is linked to concept of consensus in a field because if more researchers provide sufficient detail for others to reproduce their results, then consensus on new knowledge can more often, and more rapidly, be established. 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 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 experimental economics studies successfully replicated 61 % and 62 % of their target studies respectively (Camerer et al., 2016, 2018).

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 1203 out of 27,271 (4 %) notebooks associated with 3467 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 recognized 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). The increasing availability of large digital datasets is pushing archaeology into unexplored areas (Bevan, 2015), inviting questions about what reproducibility means for data intensive archaeological research. For example, to what extent does concern for reproducibility extend to computational reproducibility among archaeologists?

5. 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, taking guidance from similar initiatives in other fields (e.g. Ivimey-Cook et al., 2023; Nüst and Eglen, 2021). 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 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 at the time of writing).

At the time of writing (January 2025) we have completed 47 reproducibility reviews of 25 manuscripts submitted to JAS (most papers required multiple reviews). 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) offers another perspective. These 97 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.). Under this broader scope, the success rate is 7 %, a result also found in a study of 497 papers in 9 ecology journals (Kellner et al., 2025). By any measure, 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.

Fig. 5 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.

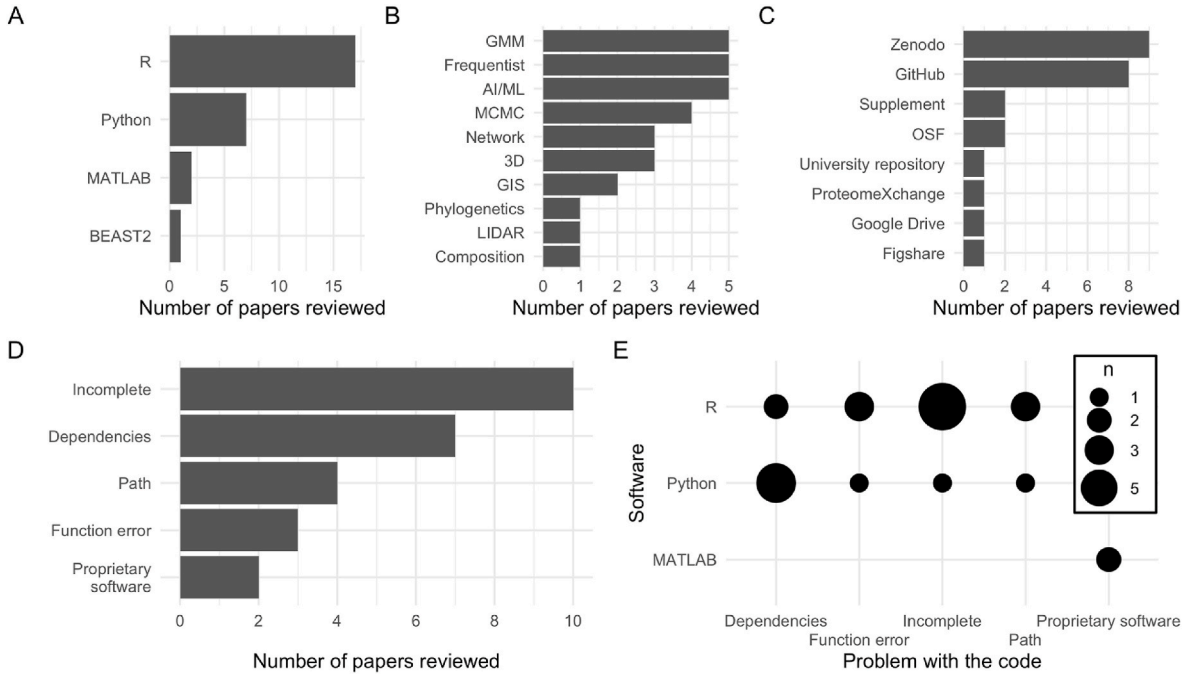

Fig. 5. Summary of reproducibility reviews for JAS. A: Primary software used for the computational analysis reported in a manuscript. B: Computational or statistical method used by the authors (GMM = geometric morphometrics; Frequentist = hypothesis tests such as chi-square and ANOVA; AI/ML = artificial intelligence and machine learning, including neural networks and deep learning; MCMC = Markov Chain Monte Carlo, i.e. Bayesian models and other simulations; Network = statistical 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.

Panels D and E of Fig. 5 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 be 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 specialised 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 the 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.

6. 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. Authors should provide their materials organised such that a reader can successfully run all code as-is, without making any manual modifications (e.g. use relative rather than absolute file paths so that readers don't have to rename or move files around to make the code work) (Sandve et al., 2013). Code and data files should be in the simplest format possible, for example a plain text R script file is smaller and easier to use than a PDF or Word document that includes R code. Script files should have the order in which they are to be run explicit in the file name, e.g. 001-load-data.R, 002-clean-data.R, 003-analyse-data.R. There are many excellent, simple, and widely-used project templates that authors can choose from that make it easy for authors to follow best practices of project organisation, e.g. Marwick et al. (2018), Figueiredo et al. (2022), Greenfeld and Community (2023), Cooper and Hsing (2017) and Wilson et al. (2017).

Second, authors should include in their compendium a README document that describes to readers the folders and files contained in the project (Abdill et al., 2024). 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 that cannot be included in the compendium due to ethical or other reasons), the key software packages needed and the version numbers that the authors used, and if the analysis takes more than a few minutes to run on a typical laptop, the hardware resources and compute time used by the author.

Related to the basic documentation provided by the README, authors should document clear, direct and obvious connections between their code and the results they present in their paper (Sandve et al., 2013). 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.

Documentation is also a key tool in tackling the problems with dependencies described in the previous section. 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., 2016, 2019; 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 (Ushey and Wickham, 2025), see examples in Vernon and Ortman (2024) and Ragno (2024), and conda and poetry for Python (AnacondaInc., 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 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. Most of our reproducibility reviews include a recommendation that the authors include a Dockerfile to manage complex dependencies efficiently.

Finally, 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 (Abdill et al., 2024; Roth et al., 2025). This ensures the project is self-contained and portable and will help the authors detect and solve issues relating to path and function errors before they submit their work for review. 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).

Fig. 6 summarises the key recommendations discussed in this section in a format that can be used as a check-list for authors submitting research for publication in JAS and other journals that do reproducibility reviews. This checklist is based on both the results presented in Fig. 5 and similar lists used by other journals, for example by the *Biometrical Journal* (Hornung et al., n.d.) and *The Review of Financial Studies* (Pérignon et al., 2024).

7. What about qualitative archaeological research?

Although the introduction of reproducibility reviews signifies a growth in computational archaeology and a desire to evaluate the research products beyond the journal article, a very substantial amount

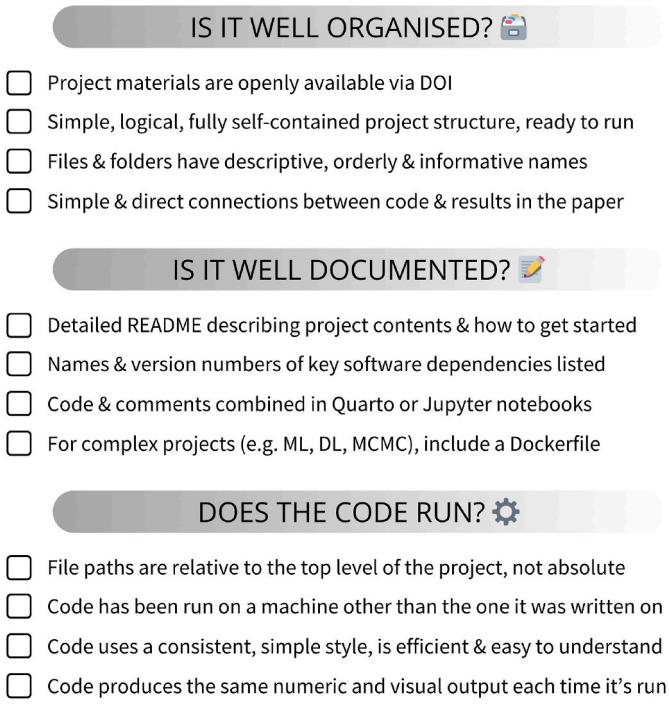

Fig. 6. Checklist summarising a small set of some of the simplest tasks authors can do that have high potential to increase reproducibility. ML = Machine Learning, DL = Deep Learning, MCMC = Monte Carlo Markov Chain simulations.

of archaeological research is qualitative, with few or no numerical data involved in making knowledge claims about the human past. For example, many archaeological questions can be answered by the simple presence or absence of artefacts or features, or qualitative comparisons of basic artefact characteristics such as shape, colour, surface treatments, and raw material. Chaîne opératoire analyses by ceramic and lithic specialists is an especially productive area of archaeological research that often relies on comparison of narratives of manufacturing processes. While these studies unquestionably count as science, because they are a systematic, empirical, and rigorous process of inquiry, should they be held to standards of reproducibility in the same way that computationally intensive research is?

A similar debate has been unfolding more broadly about the humanities where Peels and Bouter (2018a, 2018b) have argued that humanities disciplines that use empirical methods should be assessed by how well reanalysis of the original or new data using original or new methods produces the original or equivalent results. Resisting this proposal, Rijcke and Penders (2018) argue that humanities research is unique because it pursues value and meaning, and a given study can produce multiple valid answers relating to the value and meaning of a study object, so replication is irrelevant as a mark of quality. Peels (2018) disputes this uniqueness, claiming that the humanities has the same epistemic values as the sciences, however some values have more weight in the humanities while others have more weight in the sciences, and unlike in the sciences, humanities scholars often study these epistemic values themselves. A consensus seems to be emerging that for some but not all, studies in the humanities, replication is both possible and desirable, and that replication studies will differ from field to field and might even differ among various studies within a specific field (Bouter, 2019; Holbrook et al., 2019).

A key distinction here is that the locus of evaluation is empirical rather than computational, that is, concerned with appropriate reporting standards and documentation associated with physical evidence (e.g. artefacts and archives) (Stodden, 2015). A second important difference is that the debate about qualitative and humanities research is oriented towards *replication* (new data and/or new methods in an independent study to produce the same findings as original publication) rather than *reproducibility* (same data and same methods, e.g. computer code, to produce the same results as original publication) (cf Barba, 2018). This orientation to replication emphasizes triangulation techniques that compare and integrate results coming from different traditions, locations, sources and methods, which in turn supports testing whether any given inference is robust in the face of different lines of evidence (Leonelli, 2018). In sum, there are many types of qualitative and humanistic archaeology where it is possible and meaningful to maximize the chances of non-computational replication, e.g. by carefully documenting data generating processes, to produce higher quality and more impactful results.

8. 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 has additional relevanance in our current age of computationally intensive research, where my experience as a reproducibility reviewer attempting to extract useful code and data from the publication of a computational study has sometimes felt 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 Lakatos (1978)'s model 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, as indicated by the spread of journals across PC2 of Fig. 4, with distinct and sometimes non-overlapping sets of core assumptions. Some programs are more amenable to reproducibility, while others offer insights through qualitative and other 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 depicting methods 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 quantitative results.

Computational reproducibility is not a panacea; it should not be used as a universally accepted criterion for research quality (Leonelli, 2018). 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. Crema, 2025; Vaiglova, 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. This is especially the case for papers reporting results generated by long-running simulation or deep learning. These may be impracticable to fully reproduce by a typical peer reviewer who does not have access to specialised computing facilities. In a 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 (Edwards and Roy, 2017; Hagstrom, 1965). 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. This is important for updating the alignment of quantitative archaeology with normative ideals of scientific practice, such as communal sharing and organized skepticism (Merton, 1973). 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 (Dogucu, 2025; Marwick et al., 2020). A key challenge for the future is changing the dominant habitus (e.g. dispositions, skills, and ways of perceiving) of senior scholars in gatekeeping positions so that reproducibility work will be recognized and rewarded with the same level of symbolic capital afforded to novel high-impact, highly cited publications (Bourdieu, 1988).

Data availability statement

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.14897252.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Versions of this paper were presented at the Summer School on Reproducible Research In Landscape Archaeology at the Freie Universität Berlin and Christian-Albrechts-Universität zu Kiel (2017), the Big Data in Archaeology Conference at the McDonald Institute for Archaeological Research at the University of Cambridge (2019) and the Workshop on Exploring Data-Driven Solutions to Archaeological Problems at the Abu Dhabi Institute at New York University (2025). Thanks to the participants of those events for their feedback. Thanks to the JAS editors for the invitation to contribute to this special issue.

References

Abdill, R.J., Talarico, E., Grieneisen, L., 2024. A how-to guide for code sharing in biology. PLoS Biol. 22, e3002815. https://doi.org/10.1371/journal.pbio.3002815.
Alff, D., 2014. Swift's solar gourds and the rhetoric of projection. 18th Century Stud. 47, 245–260.
Allaire, J., Dervieux, C., 2024. Quarto: R Interface to 'Quarto' Markdown Publishing System.
Anaconda, Inc., 2023. Conda: a cross-platform Package and Environment Manager.
Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. Nature 533, 452–454. https://doi.org/10.1038/533452a.
Barba, L.A., 2018. Terminologies for reproducible research. https://doi.org/10.48550/arXiv.1802.03311.
Batist, Z., Roe, J., 2024. Open archaeology, open source? Collaborative practices in an emerging community of archaeological software engineers. Internet Archaeol. https://doi.org/10.11141/ia.67.13.
Becher, T., Trowler, P., 2001. Academic Tribes and Territories. McGraw-Hill, Education (UK).
Bevan, A., 2015. The data deluge. Antiquity 89, 1473–1484.
Biglan, A., 1973. The characteristics of subject matter in different academic areas. J. Appl. Psychol. 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. Am. Antiq. 28, 217–225.
Bolnick, D., Vines, T., Montgomerie, B., 2022. Ensuring data and code archive quality: why and how? (editorial). From the Editor's Desk of the American Naturalist.
Börner, K., 2010. Atlas of Science. The MIT Press.
Bourdieu, P., 1988. Homo academicus. Stanford University Press.
Bouter, L., 2019. Do the Humanities Need a Replication Drive? A Debate Rages on.
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. Rep. Subcomm. Replicability Sci. Advis. Comm. Nat. Sci. Found. Dir. Soc. Behav. Econ. Sci. 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. Nat. Hum. Behav. 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. Sci. 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. Am. Statistician 38, 261–269.
Cole, S., 1983. The hierarchy of the sciences? Am. J. Sociol. 89, 111–139. https://doi.org/10.1086/227835.
Cooper, N., Hsing, P.-Y., 2017. A Guide to Reproducible Code in Ecology and Evolution. British Ecological Society.
Crasta, S., Gannstyf, D., Developers, P., 2023. Poetry: Python Dependency Management and Packaging.
Crema, E., 2025. Statistical modelling in archaeology: some recent trends and future perspectives. J. Archaeol. Sci. 179.
Crema, E.R., Bloxam, A., Stevens, C.J., Vander Linden, M., 2024. Modelling diffusion of innovation curves using radiocarbon data. J. Archaeol. Sci. 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: Proceedings 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. Empir. Softw. Eng. 24, 381–416. https://doi.org/10.1007/s10664-017-9589-y.
Dogucu, M., 2025. Reproducibility in the classroom. Annu. Rev. Stat. Appl. 12, 89–105. https://doi.org/10.1146/annurev-statistics-112723-034436.
Editors, n.d. Reproducibility at Journal of Archaeological Science - Journal of Archaeological Science | ScienceDirect.com by elsevier.
Editors, 2012. A different agenda. Nature 487. https://doi.org/10.1038/487271a, 271–271.
Edwards, M.A., Roy, S., 2017. Academic research in the 21st century: maintaining scientific integrity in a climate of perverse incentives and hypercompetition. Environ. Eng. Sci. 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. J. Field Archaeol. 49, 391–394. https://doi.org/10.1080/00934690.2024.2391623.
Feinberg, R.A., Wainer, H., 2011. Extracting sunbeams from cucumbers. J. Comput. Graph Stat. 20, 793–810. https://doi.org/10.1198/jcgs.2011.204a.
Feinman, G., Parkinson, W., 2024. Aims and scope. J. Archaeol. Res. 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 Comput. Biol. 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. Manag. Sci. 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.
Hagstrom, W.O., 1965. The Scientific Community. Basic books, New York.
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. J. Archaeol. Sci.: 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 (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. J. Archaeol. Sci. 165, 105969. https://doi.org/10.1016/j.jas.2024.105969.
Hodder, I., 1985. Postprocessual archaeology. Adv. Archaeol. Method Theor. 1–26.
Holbrook, J.B., Penders, B., Rijcke, S. de, 2019. The humanities do not need a replication drive [WWW Document]. URL. https://www.cwts.nl/blog?article=n-r2v2a4&title=the-humanities-do-not-need-a-replication-drive.
Hornung, R., Kammer, M., Le, L., n.d. Biometrical journal - checklist for code and data supplements. Biom. J.
Isaac, G.L., 1971. Whither archaeology? Antiquity 45, 123–129. https://doi.org/10.1017/S0003598X00069283.
Ivimey-Cook, E.R., Pick, J.L., Bairos-Novak, K.R., Culina, A., Gould, E., Grainger, M., Marshall, B.M., Moreau, D., Paquet, M., Royauté, R., Sánchez-Tójar, A., Silva, I., Windecker, S.M., 2023. Implementing code review in the scientific workflow: insights from ecology and evolutionary biology. J. Evol. Biol. 36, 1347–1356. https://doi.org/10.1111/jeb.14230.
Kellner, K.F., Doser, J.W., Belant, J.L., 2025. Functional R code is rare in species distribution and abundance papers. Ecology 106, e4475. https://doi.org/10.1002/ecy.4475.
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. Int. J. Geogr. Inf. Sci. 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. Soc. Netw. Anal. Min. 10, 1–12. https://doi.org/10.1007/s13278-019-0619-1.
Lakatos, I., 1978. The Methodology of Scientific Research Programmes, Philosophical Papers. Cambridge University Press, Cambridge.
Landau, W.M., 2021. The targets r package: a dynamic make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. J. Open Source Softw. 6, 2959.
Latour, B., 1987. Science in Action: How to Follow Scientists and Engineers Through Society. Harvard university press.
Leonelli, S., 2018. Re-thinking reproducibility as a criterion for research quality. Res. Hist. Econ. Thought Methodol. 36, 129–146.
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. Comput. Sci. Eng. 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. J. Archaeol. Sci. 170, 106049. https://doi.org/10.1016/j.jas.2024.106049.
Malashichev, Y., 2017. From open access to open science. Biol. Commun. 3–5.
Martin, D.L., 2024. Editor's corner. Am. Antiq. 89, 163–164. https://doi.org/10.1017/aaq.2024.31.
Martinón-Torres, M., Killick, D., 2013. Archaeological Theories and Archaeological Sciences.
Marwick, B., 2024. Introducing the associate editor of reproducibility. Adv. Archaeol. Pract. 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. J. Archaeol. Method Theor 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). Am. 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. Adv. Archaeol. Pract. 8, 78–86. https://doi.org/10.1017/aap.2019.38.
Merton, R.K., 1973. The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago press.
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. J. Doc. 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. J. Archaeol. Sci. 169, 106021. https://doi.org/10.1016/j.jas.2024.106021.
Moreau, D., Wiebels, K., Boettiger, C., 2023. Containers for computational reproducibility. Nat. Rev. Methods Primers 3, 1–16. https://doi.org/10.1038/s43586-023-00236-9.
Muench, A., 2023. The roles of data editors in astronomy. Sci. 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. Ann. Sci. 2, 299–334.
Nüst, D., Eglen, S.J., 2021. CODECHECK: an open science initiative for the independent execution of computations underlying research articles during peer review to improve reproducibility. https://doi.org/10.12688/f1000research.51738.2.
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 Comput. Biol. 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. Am. Antiq. 88, 163–186. https://doi.org/10.1017/aaq.2023.4.
Peels, R., 2018. Epistemic values in the humanities and in the sciences. History Humanit. 3, 89–111. https://doi.org/10.1086/696304.
Peels, R., Bouter, L., 2018a. Humanities need a replication drive too. Nature 558. https://doi.org/10.1038/d41586-018-05454-w, 372–372.
Peels, R., Bouter, L., 2018b. The possibility and desirability of replication in the humanities. Palgrave Commun. 4, 1–4. https://doi.org/10.1057/s41599-018-0149-x.
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. Rev. Financ. Stud. 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. Front. Res. Metr. Anal. 2. https://doi.org/10.3389/frma.2017.00013.
Price, D., 1970. Citation measures of hard science, technology and nonscience. Commun. Sci. Eng. 3–22.
R Core Team, 2024. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
Ragno, R., 2024. Sheep and goats taxonomic abundance trends in 1st millennium CE southern Italy: multilevel bayesian modelling of NISP datasets. J. Archaeol. Sci. 171, 106068. https://doi.org/10.1016/j.jas.2024.106068.
Renfrew, C., Bahn, P., DeMarrais, E., 2024. Archaeology: Theories, Methods, and Practice, ninth ed. Thames & Hudson, New York.
Rijcke, S. de, Penders, B., 2018. Resist calls for replicability in the humanities. Nature 560. https://doi.org/10.1038/d41586-018-05845-z, 29–29.
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. J. Water Resour. Plann. Manag. 147, 01620001. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001368.
Roth, J., Duan, Y., Mahner, F.P., Kaniuth, P., Wallis, T.S.A., Hebart, M.N., 2025. Ten principles for reliable, efficient, and adaptable coding in psychology and cognitive neuroscience. Commun. Psychol. 3, 1–15. https://doi.org/10.1038/s44271-025-00236-3.
Samuel, S., Mietchen, D., 2024. Computational reproducibility of jupyter notebooks from biomedical publications. GigaScience 13, giad113. https://doi.org/10.1093/gigascience/giad113.
Sandve, G.K., Nekrutenko, A., Taylor, J., Hovig, E., 2013. Ten simple rules for reproducible computational research. PLoS Comput. Biol. 9. https://doi.org/10.1371/journal.pcbi.1003285.
Schmidt, S.C., Marwick, B., 2020. Tool-Driven Revolut. Archaeol. Sci. 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.
Skilton, P.F., 2006. A comparative study of communal practice: assessing the effects of taken-for-granted-ness on citation practice in scientific communities. Scientometrics 68, 73–96.
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. Soc. Stud. Sci. 30, 73–94. https://doi.org/10.1177/030631200030001003.
Smith, M.E., 2017. Social science and archaeological enquiry. Antiquity 91, 520–528. https://doi.org/10.15184/aqy.2017.19.
Stodden, V., 2015. Reproducing statistical results. Ann. Rev. Stat. Appl. 2, 1–19. https://doi.org/10.1146/annurev-statistics-010814-020127.
Stodden, V., Seiler, J., Ma, Z., 2018. An empirical analysis of journal policy effectiveness for computational reproducibility. Proc. Natl. Acad. Sci. 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. J. Archaeol. Sci. Scoping Future Archaeol. Sci.: Pap. Honour 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. Sci. Data 9, 60. https://doi.org/10.1038/s41597-022-01143-6.
Ushey, K., Wickham, H., 2025. Renv: Project Environments.
Vaiglova, P., 2025. How can we improve statistical training in archaeological science? J. Archaeol. Sci. 179, 106220. https://doi.org/10.1016/j.jas.2025.106220.
Vernon, K.B., Ortman, S.G., 2024. A method for defining dispersed community territories. J. Archaeol. Sci. 170, 106048. https://doi.org/10.1016/j.jas.2024.106048.
Vilhuber, L., 2019. Report by the AEA data editor. AEA Pap. Proc. 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. Sci. Rep. 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 Comput. Biol. 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. Nat. 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. J. Am. Stat. Assoc. 1–3.

Xiong, X., Cribben, I., 2023. The state of play of reproducibility in statistics: an empirical analysis. Am. 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. High. Educ. 4, 1–4.