Peer-reviewed version available at *Journal of Insect Science* **2020**, *20*, 6; doi:10.1093/jisesa/ieaa036

Jacob T. Wittman
Department of Entomology
University of Minnesota
1980 Folwell Avenue
St. Paul, MN 55102
Phone: 612-624-2751
E-mail: wittm094@umn.edu
ORCID: https://orcid.org/0000-0002-2220-5598

A guide and toolbox to replicability and open science in entomology

Jacob T. Wittman[1] and Brian H. Aukema[1]

[1]Department of Entomology, University of Minnesota, 1980 Folwell Ave, St. Paul, MN 551

Abstract

The ability to replicate scientific experiments is a cornerstone of the scientific method. Sharing ideas, workflows, data, and protocols facilitates testing the generalizability of results, increases the speed that science progresses, and enhances quality control of published work. Fields of science such as medicine, the social sciences, and the physical sciences have embraced practices designed to increase replicability. Granting agencies, for example, may require data management plans and journals may require data and code availability statements along with the deposition of data and code in publicly available repositories. While many tools commonly used in replicable workflows such as distributed version control systems (e.g. "git") or scripted programming languages for data cleaning and analysis may have a steep learning curve, their adoption can increase individual efficiency and facilitate collaborations both within entomology and across disciplines. The open science movement is developing within the discipline of entomology, but practitioners of these concepts or those desiring to work more collaboratively across disciplines may be unsure where or how to embrace these initiatives. This article is meant to introduce some of the tools entomologists can incorporate into their workflows to increase the replicability and openness of their work. We describe these tools and others, recommend additional resources for learning more about these tools, and discuss the benefits to both individuals and the scientific community and potential drawbacks associated with implementing a replicable workflow.

Keywords: reproducibility, open access, data curation, data mangement, pre-print servers

Peer-reviewed version available at Journal of Insect Science 2020, 20, 6; doi:10.1093/jisesa/ieaa036

Introduction

Fundamental to scientific inference is the axiom that scientific findings are replicable; in other words, no single study can provide conclusive evidence of scientific fact. According to recent high profile studies, a disconcerting percentage of the results from studies in the fields of medicine and social science may not be replicable, casting doubt on their veracity (Open Science Collaboration 2015, Fraser et al. 2018, Ioannidis 2018). Incentives to try to replicate results are limited, however, making it difficult to evaluate the reliability of the findings within a given field (Higginson and Munafò 2016). A putative "reproducibility crisis" (Ioannidis 2018) provides an opportunity to reflect on work within the discipline of entomology and how we can improve our implementation of the scientific method, such as practicing "open science". A European Union initiative to "foster the implementation of open science," FOSTER Plus, defines open science as "the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods" (FOSTER 2019). As such, open science and replicable science are tightly integrated.

Practicing open and replicable science can benefit individual scientists, the scientific community, and society. The common adage that "your most important collaborator is you six months from now, and past you will not answer emails" is easily addressed when adopting replicable practices, as replicable science requires thorough documentation of the processes undertaken during an experiment and analysis. Making data and code publicly available can facilitate working with current collaborators, while also making it possible for more serendipitous collaborations to arise through repository search engines. In an open science framework, the products of research move beyond the final publication by intentionally encompassing data and

code generated during the process. Availability increases the return on investment by funding agencies and increases the value of any taxpayer supported work (Molloy 2011, Piwowar et al. 2011). Publically available code and data can also increase citation rates (Piwowar et al. 2007, Uhlir and Schröder 2007, Piwowar and Vision 2013, Drachen et al. 2016, Mislan et al. 2016).The open science movement is likely to grow as disparate fields adopt these practices.

Openness and reproducibility requires an emphasis on transparency, proper documentation, and training in experimental and statistical methods. A methods section of a paper that is not completely transparent may fail to contain enough information permit true replication of the experiment. Researchers make many small decisions over the course of an experiment, known as "researcher degrees of freedom", that impact the results of the experiment or how they are interpreted (Gelman and Loken 2013, Ioannidis et al. 2014, Nakagawa and Parker 2015, Wicherts et al. 2016). *Ad hoc* decisions made during data analysis, for example, such as how to handle outliers or deciding how to deal with violation of statistical assumptions, can lead to false positives (Ihle et al. 2017, Munafò et al. 2017, Fraser et al. 2018).

The Center for Open Science, a non-profit organization dedicated to increasing openness, integrity, and reproducibility of research, worked with journals, funding agencies, and scientific societies to develop Transparency and Openness Promotion (TOP) guidelines (Nosek et al. 2015). The TOP guidelines provide modular standards that journals can adopt to encourage or require different aspects of open science, including transparency standards related to data citation, data availability, analytic methods, research materials, design and analysis, study preregistration, analysis plan preregistration, and replication. Over 5000 journals have adopted different levels of these standards as part of their publishing requirements, such as pre-registering experiments before executing them, devising a detailed data management plan, making data and

analysis code publicly available in an online repository, and submitting manuscripts to preprint servers, like *bioRxiv*, at the time of submission to a journal (Hampton et al. 2015, Michener 2015, Vale 2015, Munafò et al. 2017, Elmore 2018, Powers and Hampton 2019). Some members of the open science movement are pushing for the peer review process to become even more transparent, such as removing anonymity from peer review or requiring that reviews be publicly available alongside manuscripts (Ross-Hellauer 2017).

This special collection in the *Journal of Insect Science* is one example of how open science is becoming more prominent within the field of entomology. For researchers unfamiliar with the methods and terminology, however, open science may appear like a closed social clique. Here, we outline best practices for open science and introduce some of the tools used, along with some of the benefits of their adoption. Readers of this article may find that they are already practicing some of the recommended best practices, even though the terminology may be unique in the emerging science of Open Entomology (e.g. maintaing old versions of documents = "distributed version control"). It is our hope that this article can serve as unifying introduction to the language, tools, and processes of replication and open science, which will help facilitate learning and communicating about these processes within our field.

Replication

For the possibility of replication of a study to occur, the study must be thoroughly documented from start to finish. There are several behaviors that can affect how likely the results of a study will replicate, such as bias towards publishing positive results, sending negative or inconclusive results to the "file drawer", designing studies with low power, failing to detect data collection and entry errors through appropriate quality-control measures, reporting only positive findings of

several statistical analyses (i.e. *p*-hacking), deciding to collect more data because the results are not significant, restricting reporting to significant covariates rather than all measured covariates, treating outliers with opaque criteria, and hypothesizing after results are known (often referred to as "HARKing") (Csada et al. 1996, Borer et al. 2009, Pautasso 2010, Simmons et al. 2011, John et al. 2012, Gelman and Loken 2013, Head et al. 2015, MacCoun and Perlmutter 2015, Wicherts et al. 2016, Munafò et al. 2017, Fraser et al. 2018). It can be difficult to identify when research suffers from the aforementioned problems under current common practices. The methods section of most papers may not provide adequate detail; data are frequently unavailable and available data may not be reusable (when descriptive metadata are insuficient or errors exist). The sections that follow describe best practices to make research more replicable. Table 1 includes a selection of resources with more in depth information about each of the topics discussed below.

Data Curation

Proper data curation is vital; without proper data curation, data sets decrease in information value over time. Peripheral information is lost due to accidents, changes in file storage standards, moves, and/or the human limitations of researchers who switch careers, retire, or pass away (Michener et al. 1997). Data should have appropriate metadata attached that describe the structure of the data set, as well as synopses of how each variable was colleted. Proper metadata facilitates electronic searches for data sets of interest (Hampton et al. 2013). Borer et al. (2009) describe in detail important steps in the data curation process: descriptions should include each variable (e.g. "site_id" is the unique identifier for each site in the study), the type of data that are entered for each variable (e.g. integers, real numbers, text, etc.), allowable values (e.g. "Temperatures recorded should be between 20 and 33 °C"), and how missing or null values are recorded (e.g. N/A values are recorded as ".").

Different fields of entomology may have different standards for what type of metadata should be provided and the form in which it should be provided. For example, entomological studies that consist primarily of ecological data may wish to use the Ecological Metadata Language, described in Michener et al. (1997). Different types of data such as molecular data or genomic data frequently us their own metadata formatting conventions, many of which can be found in the list of Metadata Standards from the Digital Curation Centre (http://www.dcc.ac.uk/resources/metadata-standards/list). Using a cohesive metadata standard fosters similarity among field-specific datasets, which supports data discovery by other scientists, facilitates data acquisition and comprehension, and can further enable automatic data acquisition, cleaning, and analysis (Borer et al. 2009, Hampton et al. 2013, Michener 2015). Proper metadata makes data one produces more valuable to the broader scientific community by making it more accessible.

Data should be stored in a “tidy” format, which combines the practice of formatting data in a rectangular record format, where each variable forms a column and each observation forms a row, with relational databases, where types of observational units form separate tables that are linked by key variables, such as plot level and site level id variables (Borer et al. 2009, Wickham 2014, Hampton et al. 2015). Relational databases are discussed in the next paragraph. Often, the most efficient format to record data during the course of an experiment is not rectangular record format (also sometimes referred to as “long” format) but instead “wide” format (Tables 2 and 3, adapted from Wickham 2014). Most analytical software, however, expects that data are stored in a rectangular record format (Table 4, adapted from Wickham 2014). (Exceptions exist; for example community ecologists may often analyze species abundance data in the wide format, similar to Table 2). Researchers who record their data in the wide format may need to transform

the structure of their data from wide to long for analytical software, or vice versa. Further changes may be necessary while "cleaning" the data, such as checking the data for any obvious errors or calculating new variables as a combination or transformation of recorded variables. Before any changes are made to the data, however, researchers should save a version of this "raw", or unedited, data immediately after it has been entered. This practice of saving a raw version of the data ensures that any mistakes caught in the original data at a later date or mistakes made while cleaning the data and converting it from wide to long (aka "tidying" the data) can be easily remedied in the future.

If an experiment involves observations made on different levels (e.g. data may be recorded at different spatial scales such as multiple plots situated within a site, and multiple sites situated within a region), the data should be recorded so that each level of observational units has its own table of data and is linked to other levels of observation by a key variable (Borer et al. 2009, Hampton et al. 2015). This format of storing data in multiple tables that are linked by an identifying variable is called a relational database. For example, a researcher may study pest populations at multiple plots within an agricultural field, and at multiple agricultural fields across a region. All sampling locations within a single agricultural field will share variable measurements that are recorded at the field level, but may differ in variables recorded at the plot level. Additionally, agricultural fields within a single region will share measurements recorded at the region level, and will differ in those variables from fields within a different region. It would be redundant and increase the likelihood of mistakes if the observer recorded all field and region level variables alongside every plot observation. Eliminating redundancies in data sets stored electronically also reduces the size of the data set, which may be important if digital storage space is limited. Instead, a researcher can keep separate data tables for each level of observation:

a plot-level table, a field-level table, and a region-level table. Linking these tables with an identification variable allows the data to be combined or merged later as necessary. For example, plot-level observations can be linked to the field they were collected from by including a field ID variable, and field-level observations linked to the region they were collected from with a region ID variable. Relational database structures are not limited to data from field experiments, like in the provided example, but can be used to format any data set that exhibits such a hierarchical structure. Software or programs that help create relational databases are referred to as relational database management systems, and include software such as Microsoft Accecss or MySQL, among others. It is not necessary to use a program designed specifically as a relational database management system, however, as more general programs have the ability to manage relational databases. Microsoft Excel can be used to store data in separate tables that are linked by a key variable and scripted analytical languages, like R, have the ability to combine these tables as necessary at a later point.

When steps are taken to properly curate data and the data are made publicly available in an online repository, it increases the transparency of the data handling process and makes the data more accessible to researchers. Merging data, transforming data, or data cleaning should be performed using a scripted programming language, such as R (www.r-project.org) or Python (www.python.org). The nature of programming requires that a very clear set of instructions be delivered to the computer telling it how to clean the data. This practice leaves a clear list of actions describing how variables were transformed, how null or missing values were handled, how potential data entry errors were identified and fixed, how data were cleaned or summarized, and the order in which these actions were done. Moreover, a programming script is easy to re-run if mistakes are found later or if one wishes to incorporate more data at a later date. Most

spreadsheet software does not leave similar footprints; even with both raw data and subsequent tidy data files available, it may not be clear exactly how the tidy data version was produced. One exception is the program OpenRefine, which is freely available online (https://openrefine.org/). OpenRefine is a free "Excel-like" tool that offers traditional spreadsheet functionality while also automatically producing a reproducible script of actions taken during cleaning and analysis. Most new tools will have an associated learning curve, but in the case of R, there are hundreds, if not thousands, of resources freely available on the internet to learn the basics of working in the R environment. Novice programmers may wish to investigate R Commander, a graphical user interface for R that provides drop down menus for common commands used in R. For users interested in learning more about programming, a good place to start is the Data Carpentry website (https://datacarpentry.org), which offers free workshops and lessons in data analysis skills, or the book *R for Data Science* (Wickham and Grolemund 2016), which is available for free online (https://r4ds.had.co.nz/).

Replicable Analyses

To permit another user to replicate an analysis, it is important to document all choices made during the analysis. Any analyses performed should be curated similarly to the data. However, it is also important that careful consideration to the analysis be done prior to carrying out the experiment. This practice is formalized in the process of preregistering a study, which is a common practice in the medical field. When a study is preregistered, the study authors are asked to think critically, specifically, and exhaustively about their *a priori* hypotheses, methods, and analysis (Wicherts et al. 2016). This includes, but may not be limited to, a specific statement of the hypotheses to be tested, how the data will be analyzed, and how any issues such as violations of statistical test assumptions or how outliers will be handled. Preregistrations are submitted to a

Peer-reviewed version available at Journal of Insect Science 2020, 20, 6; doi:10.1093/jisesa/ieaa036

granting agency or posted publicly online and serve as a record to compare the final manuscript, if one is produced, to the original plan. These steps help separate exploratory analysis from confirmatory analysis. This separation reduces the temptation to hypothesize after the results are known, which can lead to bias in publications and inflated false positive rates (Munafò et al. 2017). "Researcher degrees of freedom", defined as the methodological and analytical choices made by an investigator from the start of an investigation to the end, are often not completely documented even though seemingly arbitrary decisions may influence the final presentation of results (Gelman and Loken 2013, Wicherts et al. 2016). Thoughtful layout of the statistical plan during preregistration prevents researchers from intentionally or unintentionally abusing researcher degrees of freedom.

As with data cleaning, analysis and visualization should be done using a scripted programming language. Working within a scripted program serves as a notebook to document what analyses were run, how assumptions of those analyses were tested, and is serves as a record of all results. Maintaining a scripted analysis makes it easier to update results or graphics if errors in the data or analyses are caught at a later point, without additional intellectual overhead spent on trying to remember exactly what was done.

An additional step beyond using a scripted language for data cleaning and analyses is incorporation of a distributed version control system, such as "git", that is very useful for tracking changes to programming scripts. As changes or additions are made to scripted analyses, changes are "committed" to git, which is a command that creates a history of file changes. For example, when computer code to execute a statistical analysis suddenly stops working while elaborating on the script/analysis, a researcher can easily restore the file to a commit point where the program was working. Git is also able to highlight differences between versions of a file,

which helps safeguard against accidental changes to components that should have remained static. The git program is a command-line program, requiring the user to input text commands directly into a computer terminal, which can be very intimidating to a beginner. Graphical-user interface (GUI) programs exist to make using git much easier. Such programs include GitKraken or GitHub Desktop, while R Studio has git GUI functionality built into the program. This makes integrating git into the workflow of an R user much easier. Storage is often linked to an associated cloud service, such as GitHub, GitLab, or BitBucket, making it accessible to researchers from anywhere with internet access rather than depending upon a laptop hard drive or USB thumb drive that are prone to being misplaced. The suite of office programs provided by Google (e.g. Docs, Sheets, Slides, etc.) also provide a more accessible, albeit limited, form of version control for text documents, spreadsheets, and slide shows.

Preprint Servers

A preprint server is a webpage where researchers may upload “preprints”, or early drafts, of manuscripts before they are submitted to a journal, thus allowing researchers to solicit feedback from the broader scientific community and providing immediate and open access to their results. Manuscripts submitted to a preprint server are subject to a screening process to ensure that the content is scientific, inoffensive, and not plagiarized, but the maintainers of the preprint server do not conduct any official editing or peer-reviewing functions. Instead, other users are able to provide comments on manuscripts that are uploaded. Preprint servers facilitate more public discussion and criticism of works, and allow findings to become immediately available to other researchers (Ross-Hellauer 2017, Elmore 2018). Preprint servers have their foundation in the physical sciences. The preprint server *arXiv* (pronounced “archive”) has been used by physicists

since the early 1990s to disseminate their work. The preprint server *bioRxiv* (pronounced "bio-archive") was started in 2013 to serve as a preprint repository for the biological sciences.

A manuscript on a preprint server may be updated after the authors have received feedback from the scientific community. The version of the paper is listed alongside the preprint, allowing it to serve as a record of changes between the original draft of a manuscript and the final published version. Once a version of the preprint has been published, most preprint servers allow the authors to update the original submission with a final version of the article. Authors that submit a manuscript to a preprint server retain the copyright to their work and preprint submissions are assigned a digital-object identifier (DOI), allowing them to be cited. Many journals have citation guidelines for citing preprint papers, although if the preprint has since been published in a journal it is best practice to cite the version published in the journal. The preprint website *bioRxiv* updates the preprint manuscript with a link to the published version once available. It is always good practice to check target journal submission policies before submitting a manuscript to a preprint server. Some journals disallow submission to preprint servers before submitting to the journal, while others will not allow updates to the preprint article while the article is under review at the journal.

Open Access Publishing

Open access publishing is the process of removing barriers to accessing and sharing research published in scholarly journals. Open access (OA) is often divided into two main types: Gold OA, where the publisher provides OA to the article, and Green OA, where the author(s) of the article archives the published work in an openly accessible space, such as a personal website, preprint server, or public repository (Laakso et al. 2011, Tennant et al. 2016). Gold OA often requires the author to pay extra fees associated with the open publication of their article, whereas

Green OA may be free or at low costs depending on where the article is made available. With the growth of the internet and the shift from traditional print publishing to more web-based publishing, the number of open access journals and open access journal articles has increased by 18% and 30%, from 2000 to 2009, respectively (Kaiser 2010, Laakso et al. 2011). Increases in the number of OA access articles has continued to grow since 2009, albeit at a slower rate (Piwowar et al. 2018).

The growth in OA publishing can be linked to growing beliefs that scientific research, especially publically funded research, is a public good that should be freely available (Paul et al. 2010, Grand et al. 2012, Tennant et al. 2016). Proponents of OA publishing cite an increase in public engagement (Stodden 2010), increase in public trust (Grand et al. 2012), and a decrease in inequality among countries and research institutions (Odlyzko 2006) as some of the societal benefits that result from making research more readily available. Additionally, OA articles have been shown to have higher citation rates than articles that are locked behind paywalls (Antelman 2004, Eysenbach 2006, Tennant et al. 2016, Piwowar et al. 2018). A good review of the academic, economic, and societal impacts of OA publishing can be found in Tennant (2016).

There are concerns about the cost of OA publishing. Journals may charge in excess of $1,000 USD to publish an article open access, presenting a barrier to researchers and research groups who cannot afford to pay those fees. Gold OA is more expensive than Green OA. Some institutions and funders have funds available to support OA publishing or allow researchers to expense publication fees on grants. The Public Library of Science (PLoS) maintains a partial list of such funding sources from across the global (https://plos.org/open-access-funds).

Public Repositories

To better ensure the availability of digital research materials into the future, such materials should be uploaded to a public data repository in a non-proprietary file format. Data that consists of text and numbers in rectangular record format is commonly stored in a comma-delimited file (".csv"), for example. A csv file is a basic text file that contains all the data values separated or delimited by commas. Other formats exist, such as tab- or space-delimited files. Most modern spreadsheet software is able to read basic delimited text files and convert them into a spreadsheet format that is more human-readable and works with the analytical software. Using non-proprietary formats ensures the data will be usable in the event that the most commonly-used proprietary software, like Microsoft Excel, eventually changes or disappears.

Public repositories remove the requirement for the researcher to maintain and provide research material as requested. Multiple studies have shown that researchers are not always able to provide the data associated with a manuscript and that uploading data and code to public repositories increases the likelihood that such materials are available (Leberg and Neigel 1999, Wicherts et al. 2006, Savage and Vickers 2009, Vines et al. 2013). One study looking at 516 articles in the field of ecology found that data availability decreases through time; authors of papers published 20 years ago could provide the associated data less than 50% of the time (Vines et al. 2014).

There are a variety of public repositories that will accept many types of electronic files. Many academic institutions operate their own public data repositories that can be used by associated researchers at minimal cost (i.e., even free). One popular repository not affiliated with a university is Dryad (https://datadryad.org), which specializes in life sciences data and code. Other repositories accept other types of data, such as GenBank

(www.ncbi.nlm.nih.gov/genbank/) for genetic sequence data. There may be costs associated with submitting to certain repositories, but at least one ecological journal at the time of writing, *Oikos*, has integrated their submission system with Dryad and will cover the Data Processing Charge. The journal *Scientific Data* provides a list of public data repositories across a variety of scientific disciplines (https://www.nature.com/sdata/policies/repositories).

Some data may include sensitive information that should not be made publicly available for legal or ethical reasons, such as health information, the location of endangered species, or other personal information. Steps can often be taken to deidentify data or remove the sensitive information and still post most of the data to a public repository. If the data are unusable without the sensitive information, authors should explicitly state so in a data availability statement.

Summary

Reproducible and open science is vital for ensuring inferences are valid and reliable. It may take extra care and time to ensure research is reproducible, but often these costs are outweighed by the benefits to ourselves and other researchers. The learning curve associated with incorporating these practices into the scientific workflow may be daunting, especially in light of the busy schedule already maintained by most scientists. It is our hope that this article will provide researchers with the foundation to begin incorporating these tools slowly and to seek out resources on their own, and as their schedule allows. For example, the first author of this manuscript wished to learn how to create documents in R Markdown while writing this work. This tool, provided in the software R Studio, allows users to interweave R code and prose such that users integrate their R code directly into the manuscript, choosing which code and/or outputs to display or hide in the final PDF or Word Document produced. During the writing process, this

author discovered that making tables in markdown was not trivial. As such, he opted to make them in Microsoft Word instead. The author hopes to learn this skill at a later date, as time allows.

By making tools and data more widely available, the transparency and trust in science as an institution is increased (Stodden 2010, Grand et al. 2012, Grimes et al. 2018). Additionally, we increase the value of our work when we view every step in the research cycle, not just the final manuscript, as a potential valuable product and make those products publicly available. Data, code, and the specific details of our work all provide value to the scientific community when they are easily accessible. Science serves an important role in confronting many global issues, such as resource use, invasive species, and climate change, and maintaining societal trust in the scientific process is vital to ensure that scientists remain trusted sources of information on these issues (Leiserowitz et al. 2013). Practicing open science helps ensure that science, especially publically funded science, remains transparent and accessible to the public and helps maintain public trust (Beardsley 2010).

Acknowledgements

This work was supported by USDA McIntre-Stanns project MIN-17-095 and the College of Food, Agricultural, and Natural Resource Sciences at the University of Minnesota. We thank AK Tran and S Robinson (University of Minnesota) for helpful insights that improved earlier drafts of this manuscript.

Antelman, K. 2004. Do open-access articles have a greater research impact? College & Research Libraries. 65: 372–382.

Beardsley, T. M. 2010. The biologist's burden. BioScience. 60: 483–483.

Borer, E. T., E. W. Seabloom, M. B. Jones, and M. Schildhauer. **2009**. Some simple guidelines for effective data management. Bulletin of the Ecological Society of America. 90: 205–214.

Csada, R. D., P. C. James, and R. H. M. Espie. **1996**. The "File Drawer Problem" of non-significant results: does it apply to biological research? Oikos. 76: 591–593.

Drachen, T. M., O. Ellegaard, A. V. Larsen, and S. B. F. Dorch. **2016**. Sharing data increases citations. LIBER Quarterly. 26: 67–82.

Elmore, S. A. 2018. Preprints: what role do these have in communicating scientific results? Toxicologic Pathology. 46: 364–365.

Eysenbach, G. 2006. Citation advantage of open access articles. PLoS Biology. 4: 692–698.

FOSTER. **2019**. Open Science Definition | FOSTER. https://www.fosteropenscience.eu/resources.

Fraser, H., T. Parker, S. Nakagawa, A. Barnett, and F. Fidler. **2018**. Questionable research practices in ecology and evolution. PLoS ONE. 13: e0200303.

Gelman, A., and E. Loken. **2013**. A garden of forking paths. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Grand, A., C. Wilkinson, K. Bultitude, and A. F. T. Winfield. **2012**. Open science. Science Communication. 34: 679–689.

Grimes, D. R., C. T. Bauch, and J. P. Ioannidis. **2018**. Modelling science trustworthiness under publish or perish pressure. Royal Society Open Science. 5: 171511.

Hampton, S. E., S. S. Anderson, S. C. Bagby, C. Gries, X. Han, E. M. Hart, M. B. Jones, W. C. Lenhardt, A. MacDonald, W. K. Michener, J. Mudge, A. Pourmokhtarian, M. P. Schildhauer, K. H. Woo, and N. Zimmerman. **2015**. The Tao of open science for ecology. Ecosphere. 6: art120.

Hampton, S. E., C. A. Strasser, J. J. Tewksbury, W. K. Gram, A. E. Budden, A. L. Batcheller, C. S. Duke, and J. H. Porter. **2013**. Big data and the future of ecology. Frontiers in Ecology and the Environment. 11: 156–162.

Head, M. L., L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions. **2015**. The extent and consequences of p-hacking in science. PLoS Biology. 13: e1002106.

Higginson, A. D., and M. R. Munafò. **2016**. Current incentives for scientists lead to underpowered studies with erroneous conclusions. PLoS Biology. 14: e2000995.

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

Ioannidis, J. P. 2018. Why most published research findings are false. PLoS Medicine. 2: 2–8.

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

John, L. K., G. Loewenstein, and D. Prelec. **2012**. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological science. 23: 524–32.

Kaiser, J. 2010. Free journals grow amid ongoing debate. Science. 329: 896–898.

Laakso, M., P. Welling, H. Bukvova, L. Nyman, B.-C. Björk, and T. Hedlund. **2011**. The development of open access journal publishing from 1993 to 2009. PLoS ONE. 6: e20961.

Leberg, P. L., and J. E. Neigel. **1999**. Enhancing the retrievability of population genetic survey data? An assessment of animal mitochondrial DNA studies. Evolution. 53: 1961–1965.

Leiserowitz, A. A., E. W. Maibach, C. Roser-Renouf, N. Smith, and E. Dawson. **2013**. Climategate, public opinion, and the loss of trust. American Behavioral Scientist. 57: 818–837.

MacCoun, R., and S. Perlmutter. **2015**. Blind analysis: Hide results to seek the truth. Nature. 526: 187–189.

Michener, W. K. 2015. Ecological data sharing. Ecological Informatics. 29: 33–44.

Michener, W. K., J. W. Brunt, J. J. Helly, T. B. Kirchner, and S. G. Stafford. **1997**. Nongeospatial metadata for the ecological sciences. Ecological Applications. 7: 330–342.

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

Molloy, J. C. 2011. The open knowledge foundation: Open data means better science. PLoS Biology. 9: e1001195.

Munafò, M. R., B. A. Nosek, D. V. Bishop, K. S. Button, C. D. Chambers, N. Percie Du Sert, U. Simonsohn, E. J. Wagenmakers, J. J. Ware, and J. P. Ioannidis. **2017**. A manifesto for reproducible science. Nature Human Behaviour. 1: 0021.

Nakagawa, S., and T. H. Parker. **2015**. Replicating research in ecology and evolution: Feasibility, incentives, and the cost-benefit conundrum. BMC Biology. 13: 88.

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. A. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. L. Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni. **2015**. Promoting an open research culture. Science. 348: 1422–1425.

Odlyzko, A. 2006. Economic costs of toll access, pp. 39–43. *In* Open Access: Key Strategic, Technical and Economic Aspects. Elsevier Ltd.

Open Science Collaboration. **2015**. Estimating the reproducibility of psychological science. Science. 349: aac4716–aac4716.

Paul, N., J. J. O'Donnell, A. Okersonz, and C. B. Taylor. **2010**. Editorial: Improving access to research. Science. 327: 393.

Pautasso, M. 2010. Worsening file-drawer problem in the abstracts of natural, medical and social science databases. Scientometrics. 85: 193–202.

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

Piwowar, H. A., and T. J. Vision. **2013**. Data reuse and the open data citation advantage. PeerJ. 2013: e175.

Piwowar, H. A., T. J. Vision, and M. C. Whitlock. **2011**. Data archiving is a good investment. Nature. 473: 285.

Piwowar, H., J. Priem, V. Larivière, J. P. Alperin, L. Matthias, B. Norlander, A. Farley, J. West, and S. Haustein. **2018**. The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. PeerJ. 2018.

Powers, S. M., and S. E. Hampton. **2019**. Open science, reproducibility, and transparency in ecology. Ecological Applications. 29: e01822.

Ross-Hellauer, T. 2017. What is open peer review? A systematic review. F1000Research. 6: 588.

Savage, C. J., and A. J. Vickers. **2009**. Empirical study of data sharing by authors publishing in PLoS journals. PLoS ONE. 4: e7078.

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

Stodden, V. 2010. Open science: Policy implications for the evolving phenomenon of user-led scientific innovation. Journal of Science Communication. 9: 1–8.

Tennant, J. P., F. Waldner, D. C. Jacques, P. Masuzzo, L. B. Collister, and C. H. Hartgerink. **2016**. The academic, economic, and societal impacts of Open Access: an evidence-based review. F1000. 3: 632.

Uhlir, P. F., and P. Schröder. **2007**. Open data for global science. Data Science Journal. 6: OD36–OD53.

Vale, R. D. 2015. Accelerating scientific publication in biology. Proceedings of the National Academy of Sciences of the United States of America. 112: 13439–13446.

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

Vines, T. H., R. L. Andrew, D. G. Bock, M. T. Franklin, K. J. Gilbert, N. C. Kane, J.-S. Moore, B. T. Moyers, S. Renaut, D. J. Rennison, T. Veen, and S. Yeaman. **2013**. Mandated data archiving greatly improves access to research data. The FASEB Journal. 27: 1304–1308.

Wicherts, J. M., D. Borsboom, J. Kats, and D. Molenaar. **2006**. The poor availability of psychological research data for reanalysis. American Psychologist. 61: 726–728.

Wicherts, J. M., C. L. Veldkamp, H. E. Augusteijn, M. Bakker, R. C. van Aert, and M. A. van Assen. **2016**. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid P-hacking. Frontiers in Psychology. 7: 1832.

Wickham, H. 2014. Tidy data. Journal of Statistical Software. 59: 1–23.

Wickham, H., and G. Grolemund. **2016**. R For Data Science, 1st ed. O'Reilly Media, Sebastopol, CA.

Tables

Table 1. A selection of resources to learn more about reproducibility and open science.

Topic	Resource name	Information	URL
Open Science	Open Science Foundation	A suite of information and resources about most aspects of open science.	https://cos.io/
Open Science	FOSTER	FOSTER is an e-learning platform with a variety of educational resources about open science.	https://www.fosteropenscience.eu/about
Metadata standards	Digital Curation Centre – What are Metadata Standards?	An introduction to metadata standards.	http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards
Metadata standards	Data curation	A list of different metadata standards with links to learn more.	www.dcc.ac.uk/resources/metadata-standards/list
Data curation and replicable analysis	Tidyverse	A page with resources to learn more about the R package "tidyverse".	https://www.tidyverse.org/learn/
Data curation and replicable analysis	Data Carpentry	An organization whose mission to "train researchers in the core data skills for efficient, shareable, and reproducible research practices." Their website has educational resources available for free covering topics including relational databases, data cleaning, and data analysis and visualization.	https://datacarpentry.org/
Replicable analysis	*R for Data Science*	An introduction to the R programming language and how to work with data in R. Available for free digitally online or a hardcopy can be purchased.	https://r4ds.had.co.nz/
Replicable analysis	Happy R with Git	An online tutorial to introduce users to using both git and GitHub, as well as working with them within R.	https://happygitwithr.com/
Public data repositories	Data repositories recommended by	A list of public data repositories that meet requirements for data	https://www.nature.com/sdata/policies/repositori

	the journal *Scientific Data*	access, preservation, and stability.	es
General data science	DataSciGuide	A searchable collection of resources maintained by data scientist Renee Teate. Available resources cover a wide variety of topics related to working with data electronically, including educational resources for different software, programming languages, statistical analyses, data storage, and data visualization.	http://www.datasciguide.com/

Table 2. A possible way to record species counts from two different sites. In this table, the site variables are the columns and the species variables represent the row. Each cell with a number is a different observation.

	Site A	Site B
Species 1	14	19
Species 2	29	46
Species 3	11	45

Table 3. Another way to record species counts from two different sites. The rows and columns from Table 1 are reversed.

	Species 1	Species 2	Species 3
Site A	14	29	11
Site B	19	46	45

Table 4. A tidy way to record species counts from two different sites, also commonly referred to as rectangular record format. Each variable is represented by a column and each observation by a row.

Site	Species	Count
Site A	Species 1	14
Site B	Species 1	19
Site A	Species 2	29
Site B	Species 2	46
Site A	Species 3	11
Site B	Species 3	45