

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. Wittman¹ and Brian H. Aukema¹

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

2 **Abstract**

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

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

21 **Introduction**

22 Fundamental to scientific inference is the axiom that scientific findings are replicable; in other
23 words, no single study can provide conclusive evidence of scientific fact. According to recent
24 high profile studies, a disconcerting percentage of the results from studies in the fields of
25 medicine and social science may not be replicable, casting doubt on their veracity (Open Science
26 Collaboration 2015, Fraser et al. 2018, Ioannidis 2018). Incentives to try to replicate results are
27 limited, however, making it difficult to evaluate the reliability of the findings within a given field
28 (Higginson and Munafò 2016). A putative “reproducibility crisis” (Ioannidis 2018) provides an
29 opportunity to reflect on work within the discipline of entomology and how we can improve our
30 implementation of the scientific method, such as practicing “open science”. A European Union
31 initiative to “foster the implementation of open science,” FOSTER Plus, defines open science as
32 “the practice of science in such a way that others can collaborate and contribute, where research
33 data, lab notes and other research processes are freely available, under terms that enable reuse,
34 redistribution and reproduction of the research and its underlying data and methods” (FOSTER
35 2019). As such, open science and replicable science are tightly integrated.

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

44 code generated during the process. Availability increases the return on investment by funding
45 agencies and increases the value of any taxpayer supported work (Molloy 2011, Piwowar et al.
46 2011). Publically available code and data can also increase citation rates (Piwowar et al. 2007,
47 Uhlir and Schröder 2007, Piwowar and Vision 2013, Drachen et al. 2016, Mislán et al.
48 2016). The open science movement is likely to grow as disparate fields adopt these practices.

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

58 The Center for Open Science, a non-profit organization dedicated to increasing openness,
59 integrity, and reproducibility of research, worked with journals, funding agencies, and scientific
60 societies to develop Transparency and Openness Promotion (TOP) guidelines (Nosek et al.
61 2015). The TOP guidelines provide modular standards that journals can adopt to encourage or
62 require different aspects of open science, including transparency standards related to data
63 citation, data availability, analytic methods, research materials, design and analysis, study
64 preregistration, analysis plan preregistration, and replication. Over 5000 journals have adopted
65 different levels of these standards as part of their publishing requirements, such as pre-registering
66 experiments before executing them, devising a detailed data management plan, making data and

67 analysis code publicly available in an online repository, and submitting manuscripts to preprint
68 servers, like *bioRxiv*, at the time of submission to a journal (Hampton et al. 2015, Michener
69 2015, Vale 2015, Munafò et al. 2017, Elmore 2018, Powers and Hampton 2019). Some members
70 of the open science movement are pushing for the peer review process to become even more
71 transparent, such as removing anonymity from peer review or requiring that reviews be publicly
72 available alongside manuscripts (Ross-Hellauer 2017).

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

83 **Replication**

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

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

100 *Data Curation*

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

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

124 Data should be stored in a “tidy” format, which combines the practice of formatting data in a
125 rectangular record format, where each variable forms a column and each observation forms a
126 row, with relational databases, where types of observational units form separate tables that are
127 linked by key variables, such as plot level and site level id variables (Borer et al. 2009, Wickham
128 2014, Hampton et al. 2015). Relational databases are discussed in the next paragraph. Often, the
129 most efficient format to record data during the course of an experiment is not rectangular record
130 format (also sometimes referred to as “long” format) but instead “wide” format (Tables 2 and 3,
131 adapted from Wickham 2014). Most analytical software, however, expects that data are stored in
132 a rectangular record format (Table 4, adapted from Wickham 2014). (Exceptions exist; for
133 example community ecologists may often analyze species abundance data in the wide format,
134 similar to Table 2). Researchers who record their data in the wide format may need to transform

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

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

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

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

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

194 *Replicable Analyses*

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

204 granting agency or posted publicly online and serve as a record to compare the final manuscript,
205 if one is produced, to the original plan. These steps help separate exploratory analysis from
206 confirmatory analysis. This separation reduces the temptation to hypothesize after the results are
207 known, which can lead to bias in publications and inflated false positive rates (Munafò et al.
208 2017). “Researcher degrees of freedom”, defined as the methodological and analytical choices
209 made by an investigator from the start of an investigation to the end, are often not completely
210 documented even though seemingly arbitrary decisions may influence the final presentation of
211 results (Gelman and Loken 2013, Wicherts et al. 2016). Thoughtful layout of the statistical plan
212 during preregistration prevents researchers from intentionally or unintentionally abusing
213 researcher degrees of freedom.

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

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

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

238 *Preprint Servers*

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

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

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

265 *Open Access Publishing*

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

272 Green OA may be free or at low costs depending on where the article is made available. With the
273 growth of the internet and the shift from traditional print publishing to more web-based
274 publishing, the number of open access journals and open access journal articles has increased by
275 18% and 30%, from 2000 to 2009, respectively (Kaiser 2010, Laakso et al. 2011). Increases in
276 the number of OA access articles has continued to grow since 2009, albeit at a slower rate
277 (Piwowar et al. 2018).

278 The growth in OA publishing can be linked to growing beliefs that scientific research, especially
279 publically funded research, is a public good that should be freely available (Paul et al. 2010,
280 Grand et al. 2012, Tennant et al. 2016). Proponents of OA publishing cite an increase in public
281 engagement (Stodden 2010), increase in public trust (Grand et al. 2012), and a decrease in
282 inequality among countries and research institutions (Odlyzko 2006) as some of the societal
283 benefits that result from making research more readily available. Additionally, OA articles have
284 been shown to have higher citation rates than articles that are locked behind paywalls (Antelman
285 2004, Eysenbach 2006, Tennant et al. 2016, Piwowar et al. 2018). A good review of the
286 academic, economic, and societal impacts of OA publishing can be found in Tennant (2016).

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

293 *Public Repositories*

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

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

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

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

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

326 **Summary**

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

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

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

352 **Acknowledgements**

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

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

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

- 360 **Borer, E. T., E. W. Seabloom, M. B. Jones, and M. Schildhauer. 2009.** Some simple
361 guidelines for effective data management. *Bulletin of the Ecological Society of America.*
362 90: 205–214.
- 363 **Csada, R. D., P. C. James, and R. H. M. Espie. 1996.** The "File Drawer Problem" of non-
364 significant results: does it apply to biological research? *Oikos.* 76: 591–593.
- 365 **Drachen, T. M., O. Ellegaard, A. V. Larsen, and S. B. F. Dorch. 2016.** Sharing data increases
366 citations. *LIBER Quarterly.* 26: 67–82.
- 367 **Elmore, S. A. 2018.** Preprints: what role do these have in communicating scientific results?
368 *Toxicologic Pathology.* 46: 364–365.
- 369 **Eysenbach, G. 2006.** Citation advantage of open access articles. *PLoS Biology.* 4: 692–698.
- 370 **FOSTER. 2019.** Open Science Definition | FOSTER.
371 <https://www.fosteropenscience.eu/resources>.
- 372 **Fraser, H., T. Parker, S. Nakagawa, A. Barnett, and F. Fidler. 2018.** Questionable research
373 practices in ecology and evolution. *PLoS ONE.* 13: e0200303.
- 374 **Gelman, A., and E. Loken. 2013.** A garden of forking paths.
375 http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.
- 376 **Grand, A., C. Wilkinson, K. Bultitude, and A. F. T. Winfield. 2012.** Open science. *Science*
377 *Communication.* 34: 679–689.
- 378 **Grimes, D. R., C. T. Bauch, and J. P. Ioannidis. 2018.** Modelling science trustworthiness
379 under publish or perish pressure. *Royal Society Open Science.* 5: 171511.
- 380 **Hampton, S. E., S. S. Anderson, S. C. Bagby, C. Gries, X. Han, E. M. Hart, M. B. Jones, W.**
381 **C. Lenhardt, A. MacDonald, W. K. Michener, J. Mudge, A. Pourmokhtarian, M. P.**
382 **Schildhauer, K. H. Woo, and N. Zimmerman. 2015.** The Tao of open science for
383 ecology. *Ecosphere.* 6: art120.
- 384 **Hampton, S. E., C. A. Strasser, J. J. Tewksbury, W. K. Gram, A. E. Budden, A. L.**
385 **Batcheller, C. S. Duke, and J. H. Porter. 2013.** Big data and the future of ecology.
386 *Frontiers in Ecology and the Environment.* 11: 156–162.
- 387 **Head, M. L., L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions. 2015.** The extent and
388 consequences of p-hacking in science. *PLoS Biology.* 13: e1002106.
- 389 **Higginson, A. D., and M. R. Munafò. 2016.** Current incentives for scientists lead to
390 underpowered studies with erroneous conclusions. *PLoS Biology.* 14: e2000995.
- 391 **Ihle, M., I. S. Winney, A. Krystalli, and M. Croucher. 2017.** Striving for transparent and
392 credible research: Practical guidelines for behavioral ecologists. *Behavioral Ecology.* 28:
393 348–354.
- 394 **Ioannidis, J. P. 2018.** Why most published research findings are false. *PLoS Medicine.* 2: 2–8.

- 395 **Ioannidis, J. P., S. Greenland, M. A. Hlatky, M. J. Khoury, M. R. Macleod, D. Moher, K. F.**
396 **Schulz, and R. Tibshirani. 2014.** Increasing value and reducing waste in research
397 design, conduct, and analysis. *The Lancet*. 383: 166–175.
- 398 **John, L. K., G. Loewenstein, and D. Prelec. 2012.** Measuring the prevalence of questionable
399 research practices with incentives for truth telling. *Psychological science*. 23: 524–32.
- 400 **Kaiser, J. 2010.** Free journals grow amid ongoing debate. *Science*. 329: 896–898.
- 401 **Laakso, M., P. Welling, H. Bukvova, L. Nyman, B.-C. Björk, and T. Hedlund. 2011.** The
402 development of open access journal publishing from 1993 to 2009. *PLoS ONE*. 6:
403 e20961.
- 404 **Leberg, P. L., and J. E. Neigel. 1999.** Enhancing the retrievability of population genetic survey
405 data? An assessment of animal mitochondrial DNA studies. *Evolution*. 53: 1961–1965.
- 406 **Leiserowitz, A. A., E. W. Maibach, C. Roser-Renouf, N. Smith, and E. Dawson. 2013.**
407 Climategate, public opinion, and the loss of trust. *American Behavioral Scientist*. 57:
408 818–837.
- 409 **MacCoun, R., and S. Perlmutter. 2015.** Blind analysis: Hide results to seek the truth. *Nature*.
410 526: 187–189.
- 411 **Michener, W. K. 2015.** Ecological data sharing. *Ecological Informatics*. 29: 33–44.
- 412 **Michener, W. K., J. W. Brunt, J. J. Helly, T. B. Kirchner, and S. G. Stafford. 1997.**
413 Nongeospatial metadata for the ecological sciences. *Ecological Applications*. 7: 330–342.
- 414 **Mislan, K. A., J. M. Heer, and E. P. White. 2016.** Elevating the status of code in ecology.
415 *Trends in Ecology and Evolution*. 31: 4–7.
- 416 **Molloy, J. C. 2011.** The open knowledge foundation: Open data means better science. *PLoS*
417 *Biology*. 9: e1001195.
- 418 **Munafò, M. R., B. A. Nosek, D. V. Bishop, K. S. Button, C. D. Chambers, N. Percie Du**
419 **Sert, U. Simonsohn, E. J. Wagenmakers, J. J. Ware, and J. P. Ioannidis. 2017.** A
420 manifesto for reproducible science. *Nature Human Behaviour*. 1: 0021.
- 421 **Nakagawa, S., and T. H. Parker. 2015.** Replicating research in ecology and evolution:
422 Feasibility, incentives, and the cost-benefit conundrum. *BMC Biology*. 13: 88.
- 423 **Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck,**
424 **C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J.**
425 **Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J.**
426 **Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. A. Madon, N. Malhotra, E.**
427 **Mayo-Wilson, M. McNutt, E. Miguel, E. L. Paluck, U. Simonsohn, C. Soderberg, B.**
428 **A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson,**
429 **and T. Yarkoni. 2015.** Promoting an open research culture. *Science*. 348: 1422–1425.

- 430 **Odlyzko, A. 2006.** Economic costs of toll access, pp. 39–43. *In* Open Access: Key Strategic,
431 Technical and Economic Aspects. Elsevier Ltd.
- 432 **Open Science Collaboration. 2015.** Estimating the reproducibility of psychological science.
433 Science. 349: aac4716–aac4716.
- 434 **Paul, N., J. J. O’Donnell, A. Okersonz, and C. B. Taylor. 2010.** Editorial: Improving access to
435 research. Science. 327: 393.
- 436 **Pautasso, M. 2010.** Worsening file-drawer problem in the abstracts of natural, medical and
437 social science databases. Scientometrics. 85: 193–202.
- 438 **Piwozar, H. A., R. S. Day, and D. B. Fridsma. 2007.** Sharing detailed research data is
439 associated with increased citation rate. PLoS ONE. 2: e308.
- 440 **Piwozar, H. A., and T. J. Vision. 2013.** Data reuse and the open data citation advantage. PeerJ.
441 2013: e175.
- 442 **Piwozar, H. A., T. J. Vision, and M. C. Whitlock. 2011.** Data archiving is a good investment.
443 Nature. 473: 285.
- 444 **Piwozar, H., J. Priem, V. Larivière, J. P. Alperin, L. Matthias, B. Norlander, A. Farley, J.
445 West, and S. Haustein. 2018.** The state of OA: A large-scale analysis of the prevalence
446 and impact of Open Access articles. PeerJ. 2018.
- 447 **Powers, S. M., and S. E. Hampton. 2019.** Open science, reproducibility, and transparency in
448 ecology. Ecological Applications. 29: e01822.
- 449 **Ross-Hellauer, T. 2017.** What is open peer review? A systematic review. F1000Research. 6:
450 588.
- 451 **Savage, C. J., and A. J. Vickers. 2009.** Empirical study of data sharing by authors publishing in
452 PLoS journals. PLoS ONE. 4: e7078.
- 453 **Simmons, J. P., L. D. Nelson, and U. Simonsohn. 2011.** False-positive psychology:
454 Undisclosed flexibility in data collection and analysis allows presenting anything as
455 significant. Psychological Science. 22: 1359–1366.
- 456 **Stodden, V. 2010.** Open science: Policy implications for the evolving phenomenon of user-led
457 scientific innovation. Journal of Science Communication. 9: 1–8.
- 458 **Tennant, J. P., F. Waldner, D. C. Jacques, P. Masuzzo, L. B. Collister, and C. H.
459 Hartgerink. 2016.** The academic, economic, and societal impacts of Open Access: an
460 evidence-based review. F1000. 3: 632.
- 461 **Uhlir, P. F., and P. Schröder. 2007.** Open data for global science. Data Science Journal. 6:
462 OD36–OD53.
- 463 **Vale, R. D. 2015.** Accelerating scientific publication in biology. Proceedings of the National
464 Academy of Sciences of the United States of America. 112: 13439–13446.

- 465 **Vines, T. H., A. Y. Albert, R. L. Andrew, F. Débarre, D. G. Bock, M. T. Franklin, K. J.**
466 **Gilbert, J. S. Moore, S. Renaut, and D. J. Rennison. 2014.** The availability of research
467 data declines rapidly with article age. *Current Biology*. 24: 94–97.
- 468 **Vines, T. H., R. L. Andrew, D. G. Bock, M. T. Franklin, K. J. Gilbert, N. C. Kane, J.-S.**
469 **Moore, B. T. Moyers, S. Renaut, D. J. Rennison, T. Veen, and S. Yeaman. 2013.**
470 Mandated data archiving greatly improves access to research data. *The FASEB Journal*.
471 27: 1304–1308.
- 472 **Wicherts, J. M., D. Borsboom, J. Kats, and D. Molenaar. 2006.** The poor availability of
473 psychological research data for reanalysis. *American Psychologist*. 61: 726–728.
- 474 **Wicherts, J. M., C. L. Veldkamp, H. E. Augusteijn, M. Bakker, R. C. van Aert, and M. A.**
475 **van Assen. 2016.** Degrees of freedom in planning, running, analyzing, and reporting
476 psychological studies: A checklist to avoid P-hacking. *Frontiers in Psychology*. 7: 1832.
- 477 **Wickham, H. 2014.** Tidy data. *Journal of Statistical Software*. 59: 1–23.
- 478 **Wickham, H., and G. Grolemund. 2016.** *R For Data Science*, 1st ed. O’Reilly Media,
479 Sebastopol, CA.
- 480

481 **Tables**

482 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	<i>R for Data Science</i>	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 es
 General data science DataSciGuide stability.
 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/>

483

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

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

487

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

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

490

491 Table 4. A tidy way to record species counts from two different sites, also commonly
 492 referred to as rectangular record format. Each variable is represented by a column and
 493 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

494

495