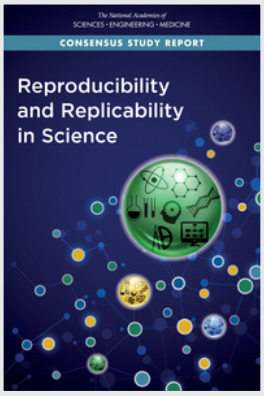

Reproducibility and Replicability in Science (2019)

DETAILS

256 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-48616-3 | DOI 10.17226/25303

CONTRIBUTORS

Committee on Reproducibility and Replicability in Science; Board on Behavioral, Cognitive, and Sensory Sciences; Committee on National Statistics; Division of Behavioral and Social Sciences and Education; Nuclear and Radiation Studies Board; Division on Earth and Life Studies; Board on Mathematical Sciences and Analytics; Committee on Applied and Theoretical Statistics; Division on Engineering and Physical Sciences; Board on Research Data and Information; Committee on Science, Engineering, Medicine, and Public Policy; Policy and Global Affairs; National Academies of Sciences, Engineering, and Medicine

SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine 2019. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.

4

Reproducibility

As defined by the committee, reproducibility relates strictly to computational reproducibility—obtaining consistent results using the same input data, computational methods, and conditions of analysis (see Chapter 3). This chapter reviews the technical and procedural challenges in ensuring reproducibility and assesses the extent of non-reproducibility in scientific and engineering research. The committee also examines factors that may deter or limit reproducibility.

WIDESPREAD USE OF COMPUTATIONAL METHODS

Most scientific disciplines today use computation as a tool (Hey et al., 2009). For example, public health researchers data mine large databases looking for patterns, earth scientists run massive simulations of complex systems to learn about geological changes in our planet, and psychologists use advanced statistical analyses to uncover subtle effects from randomized controlled experiments.

Many researchers use software at some point during their work and some are creating their own software to advance their research (Nangia and Katz, 2017). Researchers can use computation as a tool to enable data acquisition (e.g., from instruments), data management (e.g., transforming or cleaning, processing, curating, archiving), analysis (e.g., modeling,

simulation, data analysis, and data visualization), automation, and other various tasks. Computation can also be the object of study, with researchers using computing to design and test new algorithms and systems. However, the vast majority of researchers do not have formal training in software development (e.g., managing workflow processes such as maintaining code and using version control, performing unit testing).

While the abundance of data and widespread use of computation have transformed most disciplines and have enabled important scientific discoveries, this revolution is not yet reflected in how scientific results aided by computations are reported, published, and shared. Most computational experiments or analyses are discussed informally in papers, results are briefly described in table and figure captions, and the code that produced the results is seldom available. Buckheit and Donoho (1995, p. 5) paraphrase Jon Claerbout as saying, "An article about computational science [. . .] is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures."

The connection between reproducibility and transparency (i.e., open code and data) was made early by the pioneers of the reproducible research movement. Claerbout and Karrenbach (1992) advocated merging research publications with the availability of the underlying computational analysis and using a public license that allows others to reuse, copy, and redistribute the software. Buckheit and Donoho (1995, p. 4) support similar ideals, stating that "reproducibility . . . requires having the complete software environment available in other laboratories and the full source code available for inspection, modification, and application under varied parameter settings." Later, Donoho et al. (2009, p. 8) explicitly defined reproducible computational research as that in which "all details of the computation—code and data—are made conveniently available to others." The Yale Law School Roundtable on Data and Code Sharing (2010) issued a statement urging more transparency in computational sciences and offered concrete recommendations for reproducibility: assign a unique identifier to every version of the data and code, describe within each publication the computing environment used, use open licenses and nonproprietary formats, and publish under open access conditions (or post preprints). Peng (2011, p. 1226) explains:

> every computational experiment has, in theory, a detailed log of every action taken by the computer. Making these computer codes available to others provides a level of detail regarding the analysis that is greater than the analogous noncomputational experimental descriptions printed in journals using a natural language.

Nonpublic Data and Code

In many cases, sharing or submitting data and code when submitting a manuscript to a journal is the responsibility of the researcher. However, the researcher may not be allowed to do so when data or code are not publicly releasable due to licensing, privacy, or commercial reasons. For example, data or code may be proprietary as is the case often with commercial datasets; privacy laws (such as the Health Insurance Portability and Accountability Act [HIPAA]) may restrict sharing of personal information.[1]

Nonpublic data are often managed by national organizations or commercial (i.e., private) entities. In each case, protecting data and code has a reasonable goal, although one at odds with the aim of computational reproducibility. In some instances, access is allowed to researchers for both original research and reproducibility efforts (i.e., the U.S. Federal Statistical Research Data Center or the German Research Data Center of the Institute for Employment Research); in other cases, prior agreements with data or code owners will allow a researcher to share their data and code with others for reproducibility efforts (Vilhuber, 2018).

Nonpublic databases such as those storing national statistics are of particular interest to economists. Access is granted through a set of protocols. However, datasets used in research may still not be shared with others. Creation of a dataset for research is a considerable task requiring the development, in the case of databases, of queries and cleaning of the dataset prior to use. While a second researcher may have access to the same nonpublic database and the query used by the original, differences in data cleaning decisions will result in a different final dataset. Additionally, many of the large databases used by economists continuously add data so queries submitted at different times result in different initial datasets. In this case, reproducibility is not possible while replicability is (Vilhuber, 2018).

Resources and Costs of Reproducibility

Newly developed tools allow researchers to more easily follow Peng's advice by capturing detailed logs of a researchers' keystrokes or changes to code (see Chapter 6 for more details on these tools). Studies that have been designed with computational reproducibility as a key component may take advantage of these tools and efficiently track and retain relevant computational details. For studies and longstanding collaborations that have not

[1] Journals that require data to be shared generally allow some exceptions to the data sharing rule. For example, PLOS publications allow researchers to exclude data that would violate participant privacy, but they will not publish research that is based solely on proprietary data that are not made available or if data are withheld for personal reasons (e.g., future publication or patents).

designed their processes around computational reproducibility, retrofitting existing processes to capture logs of computational decisions represents a resource choice between advancing current research or redesigning a potentially large and complex system. Such studies have often developed methods for gaining confidence in the function of the system, for example, through verification and validation checks and internal reviews.

While efforts to improve reporting and reproducibility in computational sciences have expanded to the broader scientific community (Cassey and Blackburn, 2006; "Error Prone" (editorial), *Nature,* 2012; Konkol et al., 2019; Vandewalle et al., 2007), the costs and resources required to support computational reproducibility are not well established and may well be substantial. As new computational tools and data storage options become available, and as the cost of massive digital storage continues to decline, these developments will eventually make computational reproducibility more affordable, feasible, and routine.

FINDING 4-1: Most scientific and engineering research disciplines use computation as a tool. While the abundance of data and widespread use of computation have transformed many disciplines and have enabled important scientific discoveries, this revolution is not yet uniformly reflected in how scientists develop and use software and how scientific results are published and shared.

FINDING 4-2: When results are produced by complex computational processes using large volumes of data, the methods section of a traditional scientific paper is insufficient to convey the necessary information for others to reproduce the results.

RECOMMENDATION 4-1: To help ensure the reproducibility of computational results, researchers should convey clear, specific, and complete information about any computational methods and data products that support their published results in order to enable other researchers to repeat the analysis, unless such information is restricted by nonpublic data policies. That information should include the data, study methods, and computational environment:

- **the input data used in the study either in extension (e.g., a text file or a binary) or in intension (e.g., a script to generate the data), as well as intermediate results and output data for steps that are nondeterministic and cannot be reproduced in principle;**
- **a detailed description of the study methods (ideally in executable form) together with its computational steps and associated parameters; and**

- **information about the computational environment where the study was originally executed, such as operating system, hardware architecture, and library dependencies. (Library dependency,[2] in the context of research software as used here, is the relationship of pieces of software that are needed for another software to run. Problems often occur when installed software has dependencies on specific versions of other software.)**

ASSESSING REPRODUCIBILITY

When a second researcher attempts to computationally reproduce the results of another researcher's work, the attempt is considered successful if the two results are consistent. For computations, one may expect that the two results be identical (i.e., obtaining a bitwise identical numeric result). In most cases, this is a reasonable expectation, and the assessment of reproducibility is straightforward. However, there are legitimate reasons for reproduced results to differ while still being considered consistent.[3]

In some research settings, it may make sense to relax the requirement of bitwise reproducibility and settle on reproducible results within an accepted range of variation (or uncertainty). This can only be decided, however, after fully understanding the numerical-analysis issues affecting the outcomes. Researchers applying high-performance algorithms thus recognize (Diethelm, 2012) that when different runs with the same input data produce slightly different numeric outputs, each of these results is equally credible, and the output must be understood as an approximation to the correct value within a certain accepted uncertainty. Sources of the uncertainty could be, for example, floating point averaging in parallel processors (see Box 4-1) or even cosmic rays interacting with processors within a supercomputer in climate change research (see Box 4-2). In other research settings, there may be a need to reproduce the result extremely accurately, and researchers must tackle variability in computations using higher-precision arithmetic or by redesigning the algorithms (Bailey et al., 2012).

[2] This definition was corrected during copy editing between release of the prepublication version and this final, published version.

[3] As briefly mentioned in Chapter 2, reproducibility does not ensure that the results themselves are correct. If there was a mistake in the source code, and another researcher used the same code to rerun the analysis, the reproduced results would be consistent but still incorrect. However, the fact that the information was transparently shared would allow other researchers to examine the data, code, and analysis closely and possibly detect errors. For example, an attempt by an economic researcher to reproduce earlier results highlighted software errors in a statistics program used by many researchers in the field (McCullough and Vinod, 2003). Without a high level of transparency, it is difficult to know if and where a computational error may have occurred.

BOX 4-1
Parallel Processing and Numerical Precision

Although it may seem evident that running an analysis with identical inputs would result in identical outputs, this is sometimes not true. One condition under which computed results can vary between runs of the same computational analysis occurs when using computers that rely on parallel processors. Two factors are at play: the way that numbers are *represented* in a computer, and *how individual processors cooperate* in a multicore or distributed system.

Numbers are represented in a computer using floating-point representation, consisting of a number of *significant digits* scaled by an exponent in a fixed base. For example, the speed of light is 299,792,458 m/s; in normalized floating-point representation, this is 2.99792458×10^8 (in base 10). The number of significant digits gives the *precision* of the floating-point approximation. Nine digits are needed for the exact value of the speed of light, but computers store numbers with limited precision and will round this to 2.997924×10^8 when working with only seven digits of precision. If some calculation were to involve, say, adding a speed of 10 m/s to the speed of light, the rules of floating-point arithmetic mean that to add the numbers, the smaller one has to be shifted to the same exponent as the larger one, so 10 m/s is represented as 0.00000010×10^8, which with seven-digit precision gets rounded off to zero. Adding floating-point numbers of disparate scales can thus result in lost accuracy in the result.

Diethelm (2012) discusses the limits of reproducibility in high-performance (parallel) computing, given the approximate nature of floating-point arithmetic. When a large calculation (such as adding millions of numbers) is divided up so that many processors cooperate in obtaining the result in parallel, the order in which each processor finishes computing (its partial sum) cannot be guaranteed. Partial results get computed, and loss of accuracy may occur when the numbers involved have disparate scales (as described above). The final result will be different depending on the order in which the partial results are gathered together by the master process. (In mathematical terms, floating-point addition is commutative but not associative.) It is possible to prevent this lack of (numerical) reproducibility, but doing so involves artificial synchronization points in the calculation, which degrades performance. When the research requires expensive simulations that run for many days on supercomputers, the focus of research teams is understandably on maximizing performance. Thus, there is a tension between computational performance and strict numerical reproducibility of the results in parallel computing.

BOX 4-2
Reproducing Climate Model Results

For global climate models (GCMs), computational reproducibility refers to the ability to rerun a model with a given set of initial conditions and produce the same results. Such a result is achievable for short time spans and individual locations and is essential for model testing and software debugging, but the dominance of this definition as a paradigm in the field is giving way to a more statistical way of understanding model output.

Historically, climate modelers believed that they needed the more rigid definition of bitwise reproduction because the nonlinear equations governing Earth systems are chaotic and sensitive to initial conditions. However, this numerical reproducibility is difficult to achieve with the computing arrays required by modern GCMs. There is also a long history of occurrences in the models that have caused random errors and have never been reproduced, such as possible cosmic ray strikes.[a] Other reported events in uncontrolled model runs may or may not have been the result of internal model variability or software problems (see, e.g., Hall and Stouffer, 2001; Rind et al., 2018).

Reproducing the conditions that cause these random events is difficult, and scientists' lack of understanding of their effects diminishes the utility of the model. Features of computer architecture that undermine the ability to achieve bitwise reproducibility include fused multiply-add, which cannot preserve the order of operations, memory details, and issues of parallelism when a calculation is divided across multiple processors (see Box 4-1). Moreover, the environment in which GCMs are run is fragile and ephemeral on the scale of months to years, as compilers, libraries, and operating systems are continually updated, such that revisiting a 10-year-old study would require an impractical museum of supercomputers.

Retaining bitwise reproducibility will become even more difficult in the near future as machine-learning algorithms and neural networks are introduced. Therefore, scientists are also interested in representing stochasticity in the physical models by harnessing noise inherent within the electronics, and some current devices have mixed or variable bit precision.

---

[a]Cosmic ray strikes within computer hardware are another source of undetected error, and by mapping errors in model output, researchers have been able to reconstruct the path of a particle as it passed through the memory of a supercomputer stack. Therefore, the focus of the discipline has not been on model run reproducibility, but rather on replication of the model phenomena that are observed and their magnitudes (Hansen et al., 1984).

SOURCE: Adapted from Bush (2018, pp. 12-13).

A computational result may be in the form of confirming a hypothesis that entails a complex relationship among variables. Consider this example: On observing a marked seasonal migration of a species of butterflies between Europe and North Africa, researchers posed the hypothesis that the migratory strategy evolved to track the availability of host plants (for breeding) and nectar sources (Stefanescu et al., 2017). After collecting field data of plant abundance and butterfly populations, the researchers built statistical models to confirm a correlation in the temporal patterns of migration and plant abundance. The computational results were presented in the form of model parameter estimates, computed using statistical software and custom scripts. A consistent computational result, in this case, means obtaining the same model parameter estimates and measures of statistical significance within some degree of sampling variation.

Artificial intelligence and machine learning present unique new challenges to computational reproducibility, and as these fields continue to grow, the techniques and approaches for documenting and capturing the relevant parameters to enable reproducibility and confirmation of study results needs to keep pace.

FINDING 4-3: Computational reproducibility, within the range of thoughtfully assessed uncertainties, can be expected for research results given sufficient access and description of data, code, and methods, with a few notable exceptions, such as complex processing techniques and the use of proprietary or personal information.

FINDING 4-4: Understanding the limits of computational reproducibility in increasingly complex computational systems, such as artificial intelligence, high-performance computing, and deep learning, is an active area of research.

RECOMMENDATION 4-2: The National Science Foundation should consider investing in research that explores the limits of computational reproducibility in instances in which bitwise reproducibility is not reasonable in order to ensure that the meaning of consistent computational results remains in step with the development of new computational hardware, tools, and methods.

THE EXTENT OF NON-REPRODUCIBILITY

The committee was asked to assess what is known and, if necessary, identify areas that may need more information to ascertain the extent of non-reproducibility in scientific and engineering research. The committee examined current efforts to assess the extent of non-reproducibility within several fields, reviewed literature on the topic, and heard from expert panels

during its public meetings. It also drew on the previous work of committee members and other experts in the reproducibility of research. A summary of the reproducibility studies assembled by the committee is shown in Table 4-1.

As noted earlier, transparency is a prerequisite for reproducibility. Transparency represents the extent to which researchers provide sufficient information to enable others to reproduce the results. A number of studies have examined the extent of the availability of computational information within particular fields or publications as an indirect measure of computational reproducibility.

Most of the studies shown in Table 4-1 assess transparency and are thus indirect measures of computational reproducibility. Four studies listed in Table 4-1 are results of direct reproducibility (reruns of the available data and code): Dewald et al. (1986), Jacoby (2017), Moraila et al. (2013), and Chang and Li (2018). In the Dewald study, nine original research results were reproduced in a 2-year effort; of the nine, four were unsuccessful. Jacoby described the standing contract of the *American Journal for Political Science* with a university to computationally reproduce every article prior to publication; he reported to the committee that each article requires approximately 8 hours to reproduce. In Moraila's effort, software could be built for fewer than one-half of the 231 studies, highlighting the challenges of reproducing computational environments. Chang and Li were able to reproduce the results of one-half of the 67 studies they examined.

Notable in the studies listed above is the lack of a uniform standard for success or failure. The determination of transparency has layers of success. For example, downloadable data or code, downloadable data and code but not functioning, or available after a single request of the author. Similar assessments are shown for reproducibility attempts, such as the "near" successful results provided by Dewald.

FINDING 4-5: There are relatively few direct assessments of reproducibility, replaying the computations to obtain consistent results, in comparison to assessments of transparency, the availability of data and code. Direct assessments of computational reproducibility are more limited in breadth and often take much more time and resources than assessments of transparency.

CONCLUSION 4-1: Assessments of computational reproducibility take more than one form—indirect and direct—and the standards for success of each are not universal and not clear-cut. In addition, the evidence base of non-reproducibility of computations across science and engineering research is incomplete. These factors contribute to the committee's assessment that determining the extent of issues related to

TABLE 4-1 Examples of Reproducibility-Related Studies

Author	Field	Scope of Study	Reported Concerns
Prinz et al. (2011)	Biology (oncology, women's health, cardiovascular health)	Data from 67 projects within Bayer HealthCare	Published data in line with in-house results: ~20%-25% of total projects.
Iqbal et al. (2016)	Biomedical	An examination of 441 biomedical studies published between 2000 and 2014	Of 268 papers with empirical data, 267 did not include a link to a full study protocol, and none provided access to all of the raw data used in the study.
Stodden et al. (2018a)	Computational physics	An examination of the availability of artifacts for 307 articles published in the *Journal of Computational Physics*	More than one-half (50.9%) of the articles were impossible to reproduce. About 6% of the articles (17) made artifacts available in the publication itself, and about 36% discussed the artifacts (e.g., mentioned code) in the article. Of the 298 authors who were emailed with a request for artifacts, 37% did not reply, 48% replied but did not provide any artifacts, and 15% supplied some artifacts.
Stodden et al. (2018b)	Cross-disciplinary, computation-based research	A randomly selected sample of 204 computation-based articles published in *Science,* with a data-sharing requirement for publication	Fewer than one-half of the articles provided data: 24 articles had data, and an additional 65 provided some data when requested.
Chang and Li (2018)	Economics	An effort to reproduce 67 economics papers from 13 different journals	Of the 67 articles, 50% were reproduced.
Dewald et al. (1986)	Economics	A 2-year study that collected programs and data from authors who had published empirical economic research articles	Data were available for 72%-78% of the nine articles, two were reproduced successfully, three "near" successfully, and four unsuccessfully.

TABLE 4-1 Continued

Author	Field	Scope of Study	Reported Concerns
Duvendack et al. (2015)	Economics	A progress report on the number of economics journals with data-sharing requirements	In 27 of 333 economics journals, more than 50% of the articles included the authors' sharing of data and code (an increase from 4 journals in 2003).
Jacoby (2017)	Political science	A review of the results of a standing contract between *American Journal for Political Science* and universities to reproduce all articles submitted to the journal	Of the first 116 articles, 8 were reproduced on the first attempt.
Gunderson et al. (2018)	Artificial intelligence	A review of challenges and lack of reproducibility in artificial intelligence	In a survey of 400 algorithms presented in papers at two top artificial intelligence conferences in the past few years, 6% of the presenters shared the algorithm's code; 30% shared the data they tested their algorithms on; and 54% shared "pseudocode"—a limited summary of an algorithm.
Setti (2018)	Imaging	A review of the published availability of data and code for articles in *Transactions on Imaging* for 2004	For the year covered, 9% reported available code, and 33% reported available data.
Moraila et al. (2013)		An empirical study of reproducibility in computer-systems research conferences	The software could be built for less than one-half of the studies for which artifacts were available (108 of 231).

continued

TABLE 4-1 Continued

Author	Field	Scope of Study	Reported Concerns
Read et al. (2015)	Data work funded by the National Institutes of Health (NIH)	A preliminary estimate of the number and type of NIH-funded datasets; focused on those datasets that were "invisible" or not deposited in a known repository; studied published articles in 2011 cited in PubMed and deposited in PubMed Central	12% explicitly mention deposition of datasets in recognized repositories, leaving 88% (200,000 of 235,000) with invisible datasets; of the invisible datasets, approximately 87% consisted of data newly collected for the research reported, and 13% reflected reuse of existing data. More than 50% of the datasets were derived from live human or nonhuman animal subjects.
Byrne (2017)		An assessment of the open data policy of *PLOS ONE* as of 2016 (noting that rates of data and code availability are increasing)	20% of the articles have data or code in a repository; 60% of the articles have data in main text or supplemental information; and 20% have restrictions on data access.

computational reproducibility across fields or within fields of science and engineering is a massive undertaking with a low probability of success. Rather, the committee's collection of reproducibility attempts across a variety of fields allows us to note that a number of systematic efforts to reproduce computational results have failed in more than one-half of the attempts made, mainly due to insufficient detail on digital artifacts, such as data, code, and computational workflow.

Expecting computational reproducibility is considered by some to be too low of a bar for scientific research, yet our data in Table 4-1 show that many attempts to reproduce results initially fail. As noted by Peng (2016), "[Reproducibility] may initially sound like a trivial task but experience has shown that it's not always easy to achieve this seemingly minimal standard."

SOURCES OF NON-REPRODUCIBILITY

The findings and conclusion in the previous section raise a key question: What makes reproducibility so difficult to achieve? A number of factors can contribute to the lack of reproducibility in research. In addition to lack of access to nonpublic data and code, mentioned previously, the contributors include the following:

- Inadequate recordkeeping: The original researchers did not properly record the relevant digital artifacts such as protocols or steps followed to obtain the results, the details of the computational environment and software dependencies, and/or information on the archiving of all necessary data.
- Nontransparent reporting: The original researchers did not transparently report, provide open access to, or archive the relevant digital artifacts necessary for reproducibility.
- Obsolescence of the digital artifacts: Over time, the digital artifacts in the research compendium are compromised because of technological breakdown and evolution or lack of continued curation.
- Flawed attempts to reproduce others' research: The researchers who attempted to reproduce the work lacked expertise or failed to correctly follow the research protocols.
- Barriers in the culture of research: Lack of resources and incentives to adopt computationally reproducible and transparent research across fields and researchers.

The rest of this section explores each of these factors.

Inadequate Recordkeeping

The information that needs to be shared in order for research to be reproducible may vary depending on the type of research and the methods and tools used. However, the essential component is that the relevant information required to obtain a consistent result by another researcher (also referred to as the full compendium of artifacts) must be provided by the original researcher. In order to transparently report and share the full compendium of artifacts required for reproducibility, a researcher must first take care to adequately record a detailed *provenance* of all of the research results. Provenance refers to information about how a result was produced and it includes how, when, and who collected any data; what steps were followed to transform, curate, or clean them; and what software (and its version) was used to analyze them (Davidson and Freire, 2008).

In general, the computational details that need to be captured and shared for reproducible research include data, code, parameters, computational environment, and computational workflow including

- the data that were used in the analysis,[4] formatted appropriately for the research question, and complemented with standard or sufficient metadata;
- written statements in a programming language (i.e., the source code of the software used in the analysis or to generate data products) including models, data processing scripts, and software notebooks;
- numeric values of all configurable settings for software, instruments, or other hardware—that is, the parameters—for each individual experiment or run;
- detailed specification of computational environment including system software and hardware requirements, including the version number of each software used; and
- computational workflow, which is a collection of data processing scripts, statistical model specification, secondary data, and code that generated tables and figures in final published form (i.e., the computational workflow for how the software applications are configured and how the data flows between them).

Meticulous and complete recordkeeping is increasingly challenging and potentially time consuming as scientific workflows involve ever more intricate combinations of digital and physical artifacts and entail complex computational processes that combine a multitude of tools and libraries.[5] Satisfying all of these challenging conditions for transparent computation

[4] Final datasets used in analysis are the result of data collection and data culling (or cleaning). Decisions related to each step must be captured.

[5] For example, consider a scientific workflow that involves processing an image captured by an instrument, where the final presentation of the image enables the researcher to glean understanding from the data. If the researcher used image-processing software through a graphical user interface (GUI)—that is, by clicking and dragging graphical elements on the computer screen—it might be impossible for another researcher to subsequently reproduce the resulting image. For this reason, reproducibility advocates find fault with any interactive programs "unless they include the ability to arrive in any previous state by means of a script" (Fomel and Claerbout, 2009, p. 6). Some observers go as far as saying that "two technologies are enemies of reproducible research: GUI-based image manipulation, and spreadsheets" (Barba et al., 2017). The use of spreadsheet software impairs reproducibility because spreadsheets conflate input, output, code, and presentation (Stark, 2016). Spreadsheets inhibit one's ability to make a record of all steps taken to construct a full analysis of the data, and they are notoriously hard to debug. Hettrick (2017) describes the difficulties faced when trying to reproduce an analysis originally conducted on spreadsheet software, and he concluded that it is "almost impossible to reconstruct the logic behind spreadsheet-based analysis."

requires that researchers are highly motivated to ensure reproducibility. If will and incentives are lacking, it is easier for researchers to forego creating the conditions for reproducibility, as suggested by the results of reproducibility studies shown in Table 4-1. Manually keeping track of every decision in the process to include the details in a scientific paper is time-consuming and potentially error prone. Tools are available and more are being developed to autocapture relevant details in these complex environments (see Chapter 6).

Nontransparent Reporting

A second barrier to computational reproducibility is the lack of sharing or insufficient sharing of the full compendium of artifacts necessary to rerun the analysis, including the data used,[6] source code, information about the computational environment, and other digital artifacts. This information may not be reported for a number of reasons.

First, a researcher may be unaware of a norm to share the information or unaware of the details necessary to ensure reproducibility (as detailed above). Second, a researcher could be unwilling to share to ensure priority in patenting or publishing or because he or she does not see any benefit to sharing. Third, a researcher might lack the ability to share due to limited infrastructure (i.e., tools to capture the provenance or a repository to store the data or code), nonpublic restrictions (see the Nonpublic Data and Code section earlier in this chapter), or the compendium of artifacts is too large. For example, the sharing policies for *Science* offer ideas for where to share data, but they do not "suggest specific repositories or give instructions for hosting and sharing code and computational methods," and there "is no consensus regarding repositories, metadata, or computational provenance" (Stodden et al., 2018b, p. 2584).

Obsolescence of Digital Artifacts

The ability to reproduce published results can decline over time because digital artifacts can become unusable, inoperative, or unavailable due to

[6] Data quality issues also add to the complexity of identifying problems in a computational pipeline. According to J. Freire (New York University and committee member, personal communication), because people now must manage (e.g., ingest, clean, integrate, analyze) vast amounts of data, and data come from multiple sources with different levels of reliability, it is often not practical to curate the data. To extract actionable insight from data, complex computational processes are required. They are hard to assemble, and, once deployed, they can break in unforeseen ways (e.g., due to a library upgrade or a small change in the simulation code). If you have an analysis consisting of many steps, there are many ways that you could be wrong and that the data could be wrong.

technological breakdown and evolution or poor curation. This means that even if the original researcher properly recorded all of the relevant information and transparently reported it, and researchers with expertise and resources are available, reproduction attempts could still fail. Research software exists in an ecosystem of scientific libraries, system tools, and compilers. All of these are dynamic, receiving updates to improve security, fix bugs, or add features; some are no longer maintained and fail to operate with other software as the system evolves through upgrade. In the process of adding new features, a library could change how it interfaces with other software, making other code that depends on it unusable unless updated. Researchers often refer to this as "code rot." Potential solutions through archival systems have been proposed (see Chapter 6).

Flawed Attempts to Reproduce Others' Research

Just as researchers conducting original studies may make mistakes or have insufficient expertise to conduct the experiments or analysis properly, a researcher who is attempting to reproduce a result may also make mistakes or fail to follow the original protocols. Even when the original study qualifies as reproducible research, because all the relevant protocols were automated and the digital artifacts are available such that it is *capable* of being checked, another researcher without proper training and capabilities may be unable to use those artifacts.

Barriers in the Culture of Research

While interest in open science practices is growing, and many stakeholders have adopted policies or created tools to facilitate transparent sharing, the research enterprise as a whole has not adopted sharing and transparency as near-universal norms and expectations for reproducibility (National Academies of Sciences, Engineering, and Medicine, 2018).

As shown in Table 4-1, low levels of transparency are common. Currently, sharing and transparency are generally not rewarded in academic tenure and promotion systems, while the perception or reality that greater openness requires significant effort and apprehension about being scrutinized or "scooped" remain. In some disciplines and research groups, data are seen as resources that must be closely held, and it is widely believed that researchers best advance their careers by generating as many publications as possible using data before the data are shared. Shifting rewards and incentives will require thoughtful changes on the part of research institutions, working with funders and publishers (see Chapter 6).