Addressing The Scientific Reproducibility Crisis Through Educational Software
Integration

Audrey M. Bertin

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Table of Contents

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Dedication

You can have a dedication here if you wish.

```r
library(knitr)
hook_output = knit_hooks$get('output')
knit_hooks$set(output = function(x, options) {
  # this hook is used only when the linewidth option is not NULL
  if (!is.null(n <- options$linewidth)) {
    x = knitr:::split_lines(x)
    # any lines wider than n should be wrapped
    if (any(nchar(x) > n)) x = strwrap(x, width = n)
    x = paste(x, collapse = '\n')
  }
  hook_output(x, options)
})
```

Introduction

Potential sources:
  https://arxiv.org/abs/1401.3269
  https://academic.oup.com/isp/article-abstract/17/4/392/2528285
  https://berkeleysciencereview.com/2014/06/reproducible-collaborative-data-science/
  https://guides.lib.uw.edu/research/reproducibility/teaching
  https://escholarship.org/uc/item/90b2f5xh
  https://www.mitpressjournals.org/doi/full/10.1162/dint_a_00053

Chapter 1

An Introduction to Reproducibility

1.1 What Is Reproducibility?

In the field of data science, research is considered fully *reproducible* when the requisite code and data files produce identical results when run by another analyst, or more generally, when a researcher can "duplicate the results of a prior study using the same materials as were used by the original investigator" (Bollen et al. (2015)).

This term was first coined in 1992 by computer scientist Jon Claerbout, who associated it with a "software platform and set of procedures that permit the reader of a paper to see the entire processing trail from the raw data and code to figures and tables" (Claerbout & Karrenbach (1992)).

Since its inception, the concept of reproducibility has been applied across many different data-intensive fields, including epidemiology, computational biology, economics, clinical trials, and, now, the more general domain of statistical and data sciences (Goodman, Fanelli, & Ioannidis (2016)).

Reproducible research has a wide variety of benefits in the scientific community. When researchers provide the code and data used for their work in a well-organized and reproducible format, readers are more easily able to determine the veracity of any findings by following the steps from raw data to conclusions. The creators of reproducible research can also more easily receive more specific feedback (including bug fixes) on their work. Moreover, others interested in the research topic can use the code to apply the methods and ideas used in one project to their own work with minimal effort.

Although often confused, the concept of *reproducibility* is distinct from the similar idea of *replicability*: the ability of a researcher to duplicate the results of a study when following the original procedure but collecting new data. Replicability has larger-scale implications than reproducibilty; the findings of research studies can not be accepted unless a variety of other researchers come to the same conclusions through independent work.
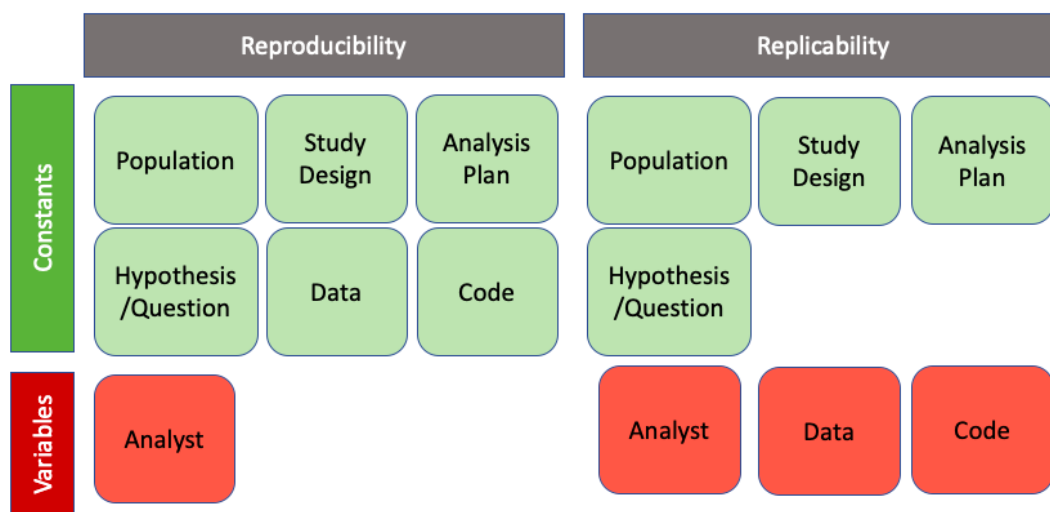
Reproducibility and replicability are both necessary to the advancement of scientific research, but they vary significantly in terms of their difficulty to achieve. Reproducibility, in theory, is somewhat simple to attain in data analyses–because code is inherently non-random (excepting applications involving random number generation) and data remain consistent, variability is highly restricted. The achievement of replicability, on the other hand, is a much more complex challenge, involving significantly more variablility and requiring high quality data, effective study design, and incredibly robust hypotheses.

1.2   The Reproducibility Crisis

Despite the relative simplicity of achieving reproducibility, a significant proportion of the work produced in the scientific community fails to meet reproducibility standards. 52% of respondents in a 2016 Nature survey believed that science was going through a "crisis" of reproducibility. Additionally, the vast majority of researchers across all fields studied reported having been unable to reproduce another researcher's results, while approximately half reported having been unable to reproduce their own (Baker (2016)). Other studies paint an even bleaker picture: a 2015 study found that over 50% of studies psychology failed reproducibility tests and research from 2012 found that figure closer to 90% in the field of cancer biology (Baker (2015), Begley & Ellis (2012)).

In the past several years, this "crisis" of reproducibility has risen toward the forefront of scientific discussion. Without reproducibility, the scientific community cannot properly verify study results.This makes it difficult to identify which information should be believed and which should not and increases the likelihood that studies sharing misleading information will be dispersed. The rise of data-driven technologies, alongside our newly founded ability to instantly share knowledge worldwide, has made reproducibility increasingly critical to the advancement of scientific understanding, necessitating the development of solutions for addressing the issue.

Academics have recognized this, and publications on the topic appear to have increased siginficantly in the last several years (Eisner (2018); Fidler & Wilcox (2018); Gosselin (2020); McArthur (2019); Wallach, Boyack, & Ioannidis (2018)).

1.3   The Components of Reproducible Research

In order to see why there is an issue with reproducibility and gain a sense of how to solve it, it is important to first understand the components of reproducibility. Essentially, to answer, "What parts does researcher need to include, or what steps do they need to take, to be able to declare their work reproducible?"

Publications attempting to answer this can be found in all areas of scientific research. However, as Goodman et al. (2016) argue, the language and conceptual framework used to describe research reproducibility varies significantly across the sciences, and there are no clear standards on reproducibility agreed upon by the scientific community as a whole.

At a minimum, according to Goodman et al. (2016), achieving reproducibility requires the sharing of data (raw or processed), relevant metadata, code, and related software. However, according to other authors, the full achievement of reproducibility may require additional components.

Kitzes, Turek, & Deniz (2017) present a collection of case studies on reproducibility practices from across the data-intensive sciences, illustrating a variety of recommendations and techniques for achieving reproducibility. Although their work does not come to a consensus on the exact standards of reproducibility that should be followed, several common trends and principles emerge from their case studies that extend beyond the minimum recommendations of Goodman et al. (2016):

1) use clear separation, labeling, and documentation in provided code,
2) automate processes when possible, and
3) design the data analysis workflow as a sequence of small steps glued together, with outputs from one step serving as inputs into the next. This is a common suggestion within the computing community, originating as part of the Unix philosophy (Gancarz (2003)).

Cooper et al. (2017) focus on data analysis completed in `R` and identify a similar list of important reproducibility components, reinforcing the need for clearly labeled, well-documented, and well-separated files. In addition, they recommend publishing a list of software dependencies and using version control to track project changes over time.

Broman (2019) reiterates the need for clear naming and file separation while sharing several additional suggestions: keep the project contained in one directory, use relative paths when accessing the file system, and include a `README` file describing the project.

The reproducibility recommendations from R OpenSci, a non-profit initiative founded in 2011 to make scientific data retrieval reproducible, share similar principles to those discussed previously. They focus on a need for a well-developed file

system, with no extraneous files and clear labeling. They also reiterate the need to note dependencies and use automation when possible, while making clear a suggestion not present in the previously-discussed literature: the need to use seeds, which allow for the saving and restoring of the random number generator state, when running code involving randomness (Martinez et al. (2018)).

Although these recommendations differ from one another, when considered in combination they provide a well-rounded picture of the components important to research reproducibility across the scientific community:

1. The basic project components are made accessible to the public:

   - Data (raw and/or processed)
   - Metadata
   - Code
   - Related Software

2. The file structure of project is well-organized:

   - Separate folders for different file types.
   - No extraneous files.
   - Minimal clutter.

3. The project is documented well:

   - Files are clearly named, preferably in a way where the order in which they should be run is clear.
   - A README is present.
   - Code contains comments.
   - Software dependencies are noted.

4. File paths used in code are not system- or user-dependent:

   - No absolute paths.
   - No paths leading to locations outside of a project's directory.
   - Only relative paths, pointing to locations within a project's directory, are permitted.

5. Randomness is accounted for:

   - If randomness is used in code, a seed must also be set.

6. Code is readable and consistently styled:

   - Though not mentioned in the sources described previously, it is also important that code be written in a coherent style. This is because code that conforms to a style guide or is written in a consistent dialect is easier to read, simplifying the process of following a researcher's work from beginning to end (Hermans & Aldewereld (2017)).

1.4   Current Attempts to Address Reproducibility in Scientific Publishing

In an attempt to increase reproducibility, leaders from academic journals around the world have taken steps to create new standards and requirements for submitted articles. These standards attempt to address the components of reproducibility listed previously, requesting that authors provide certain materials necessary for reproducing their work when they submit an article. However, these standards are highly inconsistent, varying significantly both across and within disciplines, and many only cover one or two of the six primary components, if any at all.

To illustrate this point, we will consider several case studies from journals publishing research on a variety of scientific fields.

1.4.1   Case Studies Across The Sciences

The journal whose requirements appear to align most closely with those components defined previously in Section 3 is the *American Journal of Political Science* (AJPS). In 2012, the AJPS became the first political science journal to require authors to make their data openly accessible online, and the publication has instituted stricter requirements since. AJPS now requires that authors submit the following alongside their papers (American Journal of Political Science (2016)).

- The dataset analyzed in the paper and information about its source. If the dataset has been processed, instructions for manipulating the raw data to achieve the final data must also be shared.
- Detailed, clear code necessary for reproducing all of the tables and figures in the paper.
- Documentation, including a README and codebook.
- Information about the software used to conduct the analysis, including the specific versions and packages used.

These standards are quite thorough and contain mandates for the inclusion of the vast majority of components necessary for complete reproducibility. Most journals, however, do not come close to meeting such high standards in their reproducibility statements.

For example, in the biomedical sciences, a group of editors reproesenting over 30 major journals met in 2014 to address reproducibility in their field, coming to a consensus on a set of principles they wanted to uphold (National Institutes of Health (2014)). Listed below are those relating specifically to the use of data and statistical methods:

1) Journals in the biomedical sciences should have a mechanism to check the statistical accuracy of submissions.

2) Journals should have no (or generous) limit on methods section length.

3) Journals should use a checklist to ensure the reporting of key information, including:

- The article meets nomenclature/reporting standards of the biomedical field.
- Investigators report how often each experiment was performed and whether results were substantiated by repetition under a range of conditions.
- Statistics must be fully reported in the paper (including test used, value of N, definition of center, dispersion and precision measures).
- Authors must state whether samples were randomized and how.
- Authors must state whether the experiment was blinded.
- Authors must clearly state the criteria used for exclusion of any data or subjects and must include all results, even those that do not support the main findings.

4) All datasets used in analysis must be made available on request and should be hosted on public repositories when possible. If not possible, data values should be presented in the paper or supplementary information.

5) Software sharing should be encouraged. At the minimum, authors should provide a statement describing if software is available and how to obtain it.

Even though these principles seem well-developed on the surface, they fail to meet even the basic requirements defined by Goodman et al. (2016) previously. Several of the principles are purely recommendations; there is no requirement that code be shared, nor metadata. Additionally, software requirements are quite loose, requiring no information about dependencies or software version.

We see a similar issue even in journals designed specifically for the purpose of improving scientific reproducibility. *Experimental Results*, a publication created by Cambridge University Press to address some of the reproducibility and open access issues in academia, also falls short of meeting high standards. The journal, which showcases articles from a variety of scientific decisions disciplines, states in their transparency and openness policy:

> Whenever possible authors should make evidence and resources that underpin published findings, such as data, code, and other materials, available to readers without undue barriers to access.

The inclusion of code and data are only recommended and no definition of what "other materials" may mean is provided. No components of reproducibility extending beyond those required at a minimum are even considered (Cambridge University Press (2020)).

The *American Economic Review*, the first of the top economics journals to require the inclusion of data alongside publications, has stronger guidelines than several of those mentioned previously, though not as strong as the *American Journal of Political Science*. Their Data and Code Availability Policy states the following (American Economic Association (2020)):

> It is the policy of the American Economic Association to publish papers only if the data and code used in the analysis are clearly and precisely documented, and access to the data and code is clearly and precisely documented and is non-exclusive to the authors.

These requirements are quite strict, prohibiting exceptions for papers using data or code not available to the public in the way that many other journals claiming to promote reproducibility do.

1.4.2 Case Studies In The Statistical And Data Sciences

When considering reproducibility policy, the field of Statistical and Data Sciences performs relatively well. The majority of highly ranked journals in the field contain statements on reproducibility. Some of these are quite robust, surpassing the requirements of many of the other journals discussed previously, while others are lacking.

The *Journal of the American Statistical Association* stands out as having relatively robust requirements. The publication's guideliens require that data be made publicly available at the time of publication except for reasons of security or confidentiality. It is strongly recommended that code be deposited in open repositories. If data is used in a process form, the provided code should include the necessary cleaning/preparation steps. Data must be in an easily understood form and a data dictionary should be included. Code should also be in a form that can be used and understood by others, including consistent and readable syntax and comments. Workflows involving more than one script should also contain a master script, Makefile, or other mechanism that makes it clear what each component does, in what order to run them, and what the inputs and outputs to each area (American Statistical Association (2020)).

The *Journal of Statistical Software* also has strong guidelines, though less thorough. Authors must provide *commented* source code for their software; all figures, tables, and output must be exactly reproducible on at least one platform, and random number generation must be controlled; and replication materials (typically in the form of a script) must be provided (Journal of Statistical Software (2020)).

The expectations of the *Journal of Computational and Graphical Statistics* are notably weaker, requiring only that authors "submit code and datasets as online supplements to the manuscript," with exceptions for security or confidentiality, but providing no further detail (Journal of Computational and Graphical Statistics (2020)). The *R Journal* has the same requirements, but with no exceptions on the data provision policy, stating that authors should "not use such datasets as examples" (R Journal Editors (2020)).

Perhaps the weakest reproducibility policies come from *The American Statistician* and the *Annals of Statistics*. The former appears to have no requirements, stating only that it "strongly encourages authors to submit datasets, code, other programs, and/or appendices that are directly relevant to their submitted articles," while the latter appears to have no statement on reproducibility at all (The American Statistician (2020))s.

Journal Name	Code Sharing Required?	Data Sharing Required?	Other Components Required?
Journal of the American Statistical Association	✚ ◆	✚	✚
Journal of Statistical Software	✚	✚	✚
Journal of Computational and Graphical Statistics	✚ ◆	✚ ◆	🚫
The R Journal	✚	✚	🚫
The American Statistician	🚫 ★	🚫 ★	🚫
The Annals of Statistics	🚫	🚫	🚫

★ Component recommended, but not required.
◆ Component required, but exceptions granted.

1.4.3   The Bigger Picture

The journals mentioned here are just some of the many academic publishers with reproducibility policies. While they provide a sense of the specific wording and requirements of some policies, they do not necessarily serve as a representative sample of all academic publishing. It is important to also consider the bigger picture, exploring the state of reproducibility policy in academic publishing as a whole.

Given the scale of the academic publishing network and the sheer number of journals around the world, this is not necessarily an easy task.

In order to simplify this process, academics at the Center for Open Science (COS) attempted to create a metric, called the TOP Factor. The TOP Factor reports the steps that a journal is taking to implement open science practices. It has been calculated for a wide variety of journals, though the COS is still far from scoring all of the publications that are currently available.

The TOP Factor is calculated as follows. Publications are scored on a variety of categories associated with open science and reproducibility. For each category, they receive a score between 0 (poor) and 3 (excellent) based on the degree to which they emphasize each category in their submission/publication policies. A journal's final score, which can range from 0 to 30, is the sum of the individual scores in each of the categories.

	Not Implemented	Level I	Level II	Level III
Citation standards	Journal encourages citation of data, code and materials or says nothing.	Journal describes citation of data in guidelines to authors with clear rules and examples	Article provides appropriate citation for data and materials used consistent with journal's author guidelines	Article is not published until providing appropriate citation for data and materials following journal's author guidelines
Data transparency	Journal encourages data sharing or says nothing	Article states whether data are available and, if so, where to access them	Data must be posted to a trusted repository. Exceptions must be identified at article submission	Data must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication
Code transparency	Journal encourages code sharing or says nothing	Article states whether code is available and, if so, where to access it	Code must be posted to a trusted repository. Exceptions must be identified at article submission	Code must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication
Research materials transparency	Journal encourages materials sharing or says nothing	Article states whether materials are available and, if so, where to access them	Materials must be posted to a trusted repository. Exceptions must be identified at article submission	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication

Design and analysis transparency	Journal encourages design and analysis transparency or says nothing	Journal articulates design transparency standards	Journal requires adherence to design transparency standards for review and publication	Journal requires and enforces adherence to design transparency standards for review and publication
Study preregistration	Journal says nothing	Article states whether preregistration of study exists and, if so, where to access it	Article states whether preregistration of study exists and, if so, allows journal access during peer review for verification	Journal requires preregistration of studies and provides link and badge in article to meeting requirements
Analysis plan preregistration	Journal says nothing	Article states whether preregistration of study exists and, if so, where to access it	Article states whether preregistration with analysis plan exists and, if so, allows journal access during peer review for verification	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements
Replication	Journal discourages submission of replication studies or says nothing	Journal encourages submission of replication studies	Journal encourages submission of replication studies and conducts results blind review	Journal uses registered reports as a submission option for replication studies with peer review prior to observing the study outcomes

When looking at the overall distribution of TOP Factor scores, we see a relatively grim picture: Around 50% of journals score as low as 0-5 overall, while only just over 5% score more than 15, just half of the maximum possible score. Over 40 journals failed to score a single point (Woolston (2020)).

Although it is clear that some journals have relatively strong reproducibility and openness policies, that is clearly not the norm. And many that do appear to have policies lacking in robustness, including exceptions for data privacy and security concerns or phrasing guidelines as recommendations rather than reqquirements. The field of data science stands out among the rest, with the majority of top journals having relatively robust policies.

1.4.4 Assessing the Success of Academic Reproducibility Policies

We have seen that, although not necessarily the standard, some journals across the sciences have enacted reproducibility policies. The simple implementation of a policy, however, does not ensure that its goals will be achieved. Reproducibility can only be addressed when both authors *and* journal reviewers actively implement publishing standards in practice. Without participation and dedication from all involved, reproducibility guidelines serve more as a theoretical goal than a practical achievement.

It is important to ask, then, whether academic reproducibility standards *actually* result in a greater number of reproducible publications.

Let us consider the case of the journal *Science*. *Science* instituted a reproducibility policy in 2011 and has maintained it ever since. In its original form, their policy stated the following:

> All data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of Science. All computer codes involved in the creation or analysis of data must also be available to any reader of Science. After publication, all reasonable requests for data and materials must be fulfilled. Any restrictions on the availability of data, codes, or materials...must be disclosed to the editors upon submission...

This policy is similar to many of the others considered previously, requiring the publishing of code and data with exceptions permitted when necessary.

Stodden, Seiler, & Ma (2018a) tested the efficacy of this policy in practice, emailing corresponding authors of 204 articles published in the year after *Science* first implmented its policy to request the data and code associated with their articles. The researchers only received (at least some of) the requested material from 36% of authors. This low rates were due to several factors:

- 26% of authors did not respond to email contact.
- 11% of authors were unwilling to provide the data or code without further information regarding the researchers' intentions.
- 11% asked the researchers to contact someone else and that person did not respond.
- 7% refused to share data and/or code.
- 3% directed the researchers back to their paper's supplmental information section.
- 3% of authors made a promise to follow up and then did not follow through.
- 3% of emails bounced.
- 2% gave reasons why they could not share for ethical reasons, size limitations, or some other reason.

Of the 56 papers they deemed likely reproducible, the authors randomly selected 22 and were able to replicate the results for all but 1, which failed due to its reliance on software that was no longer available.

Hardwicke et al. (2018) conducted a study on the journal *Cognition*, where researchers compared the reproducibility of published work both before and after the journal instituted an open data policy, which required that authors make relevant research data publicly available prior to publication of an article.

The researchers found a considerable increase in the proportion of data available statements (in constrast to 'data not available' statements, which could be present due to privacy or security concerns) since the implementation of the policy. Pre-open data policy, only 25% of articles had data available, while that number was a much higher 78% after the policy was put in place.

While the institution of an open data policy appears to have been associated with a significant increase in the percentage of studies with data available, further research indicates that the policy was perhaps not as effective as intended. Many of the datasets were usable in theory, but not in practice. Only 62% of the articles with data available statements had truly reusable datasets–in this case, meaning that the data were accessible, complete, and understandable. Though this is an increase from the pre-policy period, which saw 49% of articles with data availability statements as reusable in practice, it is still far from ideal.

Combining these two data points indicates that *less than half* of articles published after the open data policy was instituted actually contained truly usable data.

In this small sample of cases, we see that purely having a reproducibility statement does not necessarily mean that all, or even a majority, of published work will truly be reproducible.

1.5   Limitations on Achieving Reproducibility in Scientific Publishing

There are several reasons for this apparent divide between journal reproducibility standards and the true proportion of submitted articles that are truly reproducible. Some of these are challenges faced by the article authors, while others are faced by the journal editors.

1.5.1   Challenges for Authors

Stodden, Seiler, & Ma (2018b) conducted a survey asking over 7,700 researchers about one of the key characteristics of reproducibility – open data – and gathered information about the reasons why authors found difficulties in making their data available to the public

The main challenges listed by respondents were as follows:

- 46% identified "Organizing data in a presentable and useful way" to be difficult.
- 37% had been "Unsure about copyright and licensing."
- 33% had problems with "Not knowing which repository to use."
- 26% cited a "Lack of time to deposit data."
- 19% found the "Costs of sharing data" to be high.

The relative frequency of these issues varied across several characteristics, including author seniority, subject area, and geographical location, though authors in all categories faced some issues.

Beyond technical challenges, other reasons may lead authors to not place their focus on reproducibility. For example, some researchers might fear damage to their reputation if a reproduction attempt fails after they have provided the necessary materials.

ADD SOURCE: LUPIA, ARTHUR, AND COLIN ELMAN. (2014) Openness in Political Science: Data Access and Research Transparency. PS, Political Science & Politics 47 (1): 19–42.

Given the relatively high frequency of concern over achieving reproducibility, it follows that researchers will not make the necessary effort to do so if journal guidelines provide a way out. Policies that *recommend* the inclusion of data or that allow exceptions to open data for certain reasons are likely to be associated with a lower proportion of reproducible articles than those that make open data mandatory.

1.5.2 Challenges for Journals

In addition to the challenges faced on the part of the authors, journal reviewers face their own difficulties in ensuring reproducibility.

In order to make sure that all submitted articles comply with reproducibility guidelines, reviewers must go through them one by one and reproduce all of the results by hand using the provided materials.

This is an incredibly intensive process, as we will see in the example of the *American Journal for Political Science* (AJPS), whose reproducibility policy was discussed prevously in Chapter 1.4.1.

Jacoby, Lafferty-Hess, & Christian (2017) describe the AJPS process in detail:

Acceptance of an article for publication in the AJPS is contingent on successful reproducibility of any empirical analyses reported in the article.

After an article is submitted, staff from a third party vendor hired by AJPS go through the provided materials to ensure that they can be preserved, understood, and used by others. They then run all of the analyses in the article using the code, instructions, and data provided by the authors and compare their results to the submitted articles. Authors are then given an opportunity to resolve any issues that come up. This process is repeated until reproducibility is ensured.

Although providing a significant benefit to the scientific community, this thorough process is associated with high costs.

The verification process slows down the journal review process significantly, adding a median 53 days to the publication workflow, as many submitted articles require one or more rounds of resubmission (the average number of resubmissions is 1.7). It is also quite labor intensive, taking an average of 8 person-hours per manuscript to reproduce the analyses and prepare the materials for public release and adding significant monetary cost to AJPS.

Journals are often reluctant to take on such an intensive task due to the drastically increased burden it places on reviewers and on the publication's financial resources.

This is particularly true given that the number of submitted articles per year has been increasing over time (Leopold (2015)). Every additional submission increases the burden of achieving reproducibility, and with a large enough volume, the challenge can quickly become seemingly impossible to manage reasonably.

As a result, journals often encourage reviewers to consider authors' compliance with data sharing policies, but do not formally require that they ensure it as a criterion for acceptance (Hrynaszkiewicz (2020)).

1.6   Attempts to Address These Limitations

The previous discussion makes clear that, although reproducibility is critically important to scientific progress and academic journals are taking steps to encourage it, the scientific community is far from achieving the desired level of widespread reproducibility. In large part, this appears due to the challenge and complexity of actually achieving reproducibility. Those attempting to improve the reproducibility of work can face issues with concerns over legality of sharing data, large commitments of time or money, challenges finding a good repository, and organizing all of the many components of their work in an understandable way, among other things.

Additionally, science faces the additional challenge that many publishers do not emphasize reproducibility at all, providing many opportunities for all authors except those personally dedicated to producing reproducible work to leave reproducibility by the wayside. Many journals have no reproducibility requirements, and those that do often do not take the necessary steps to ensure that they are actually met.

These issues, however, are not impossible to overcome. Proponents of reproducibility have taken action to help address them, both through education on reproducibility and through software that helps simplify the process of achieving it.

1.6.1   Through Education

One way to address the reproducibility crisis is to educate data analysts on the topic so that they are aware of both the concept of reproducibility and how to achieve it in their own work. A natural place to focus this education is early on in the data science training pipeline as part of introductory or early-intermediate courses in undergraduate and graduate data science porgrams. (`https://arxiv.org/abs/1401.3269`) This sort of educational integration has a variety of benefits:

- Bringing reproducibility into the discussion early on gives students the tools to add knowledge to their field in the best way possible before they actually conduct any substantive analysis on their own. (Bringing the Gold Standard into the Classroom: Replication in University Teaching - Nicole Janz). This produces many long run benefits, helping lessen the burden on promoting reproducibility placed on journals and increasing the number (and percentage) of researchers doing and promoting reproducible work.

- If covered in detail as part of the data science curriculum, reproducibility will

eventually come easily to students. If learned independently, without effective
tools, reproducibility can challenging and even disheartening to try to under-
stand and succeed at. Practicing in the classroom gives students the ability
to fail without damaging their reputation, giving a great opportunity to truly
learn and understand the concepts so that they feel capable of handling them
when they begin their own research.

- The application of grading to the topic provides an incentive for students to
  pay attention, learn, and absorb the information. This same incentive does not
  exist when researchers attempt to learn about reproducibility independently. In
  that situation, internal motivation, which may be weak in some individuals, is
  the only factor present to help promote success.

Several educators, primarily at the graduate level, have realized the opportunity and
taken steps to introduce reproducibility into their courses.

The primary way of achieving this integration is through the assignment of "repli-
cation studies" in standard methods choices. In these assignments, students are given
a published study and its supporting materials and asked to reproduce the results.
The most famous course of this kind is Government 2001, taught by Gary King at
Harvard University. (Janz article) In King's course, students team up in small groups
to reproduce a previous study. To help ensure that their workflow is reproducible,
students are required to hand over their data and code to another student team who
then tries to reproduce their work once again.

In Thomas M. Carsey's intermediate statistics course at the University of North
Carolina at Chapel Hill, students must reproduce the findings of a study by re-
collecting the data from the original sources, then must extend the study by building
on the analysis.

Christopher Fariss of Penn State University asks his students to replicate a re-
search paper published in the last five years, noting that students must describe the
article and the ease in which the results replicate.

The University of California at Berkeley has a similar course to Gary King's
Harvard course, where students each take a different piece of an existing study to
work on reproducing and have to ensure that their piece fits with the piece of the
next student. (`https://berkeleysciencereview.com/2014/06/reproducible-`
`collaborative-data-science/`)

At the undergraduate level, rather than assign replication studies the way many
graduate schools tend to do, Smith College and Duke University have both inte-
grated reproducibility into their introductory courses through the requirement that
assignments be completed in the `RMarkdown` code + narration format. (`https:`
`//escholarship.org/uc/item/90b2f5xh`).

Another way to provide education on reproducibility is through the creation of
workshops that focus solely on the topic, rather than through integration as just one
part of a class.

For example, the University of Cambridge conducts a Replication Workshop,
where graduate students are asked replicate a paper in their field over eight weekly
sessions. When students encounter challenges, such as authors not responding to

queries for data or steps of the analysis being poorly defined and explained, they gain a first hand understanding of the consequences of poor transparency.

Workshops such as these are typically optional and not included as part of the primary curriculum, however, so while they may cover the topic of reproducibility in more detail than traditional courses, they often reach fewer students.

In spite of all of the advantages that these educational tools provide, "reproducibility training and assessment in data science education is largely neglected, especially among undergraduates and Master's students in professional schools..., probably because the students are usually considered to be non-research oriented." (`https://www.mitpressjournals.org/doi/pdf/10.1162/dint_a_00053`) While some examples of reproducibility education exist, they are certainly not commonplace. However, given the increased discussion and emphasis on reproducibility in academia over the past several years, it is likely that this will change, particularly if methods are provided to educators to make the integration of reproducibility into their courses simple and relatively unburdensome.

1.6.2   Through Software

Several researchers and members of the Statistical and Data Sciences community have taken action to develop software focused on reproducibility which removes some of the load on data analysts by automating reproducibility processes and checking whether certain components are achieved.

Much of this software has been written for users of the coding and data analysis language `R`. `R` is very popular in the data science community due to its open-source nature, accessibility, extensive developer and user base, and statsitical analysis-specific features.

Some of the existing software solutions are listed below:

`rrtools` (Marwick (2019)) addresses many of the issues discussed in Marwick, Boettiger, & Mullen (2018) by creating a basic reproducible structure based on the `R` package format for a data analysis project. In addition, it allows for isolation of the computer environment using `Docker`, provides a method to capture information about the versions of packages used in a project, contains tools for generating a README file, and provides an option for users to write tests to check that their functions operate as intended.

The `orderly` (FitzJohn et al. (2020)) package also focuses on file structure, requiring the user to declare a desired project structure (typically a step-by-step structure, where outputs from one step are inputs into the next) at the beginning and then creating the files necessary to achieve that structure. Its principal aim is to automate many of the basic steps involved in writing analyses, making it simple to:

1) Track all inputs into an analysis.
2) Store multiple versions of an analysis where it is repeated.
3) Track outputs of an analysis.
4) Create analyses that depend on the outputs of previous analyses.

When projects have a variety of components, `orderly` makes it easy to see inputs

and outputs change with each re-run.

`workflowr`'s (Blischak, Carbonetto, & Stephens (2019)) functionality is based around version control and making code easily available online. It works to generate a website containing time-stamped, versioned, and documented results. In addition, it manages the session and package information of each analysis and controls random number generation.

`checkers` (Ross, DeCicco, & Randhawa (2018)) allows you to create custom checks that examine different aspects of reproducibility. It also contains some pre-built checks, such as seeing if users reference packages that are less preferred to other similar ones and ensuring that the project is under version control. `renv` (Ushey & RStudio (2020)) (formerly `packrat`) helps to make projects more isolated, portable, and reproducible. It gives every project its own private package library, makes it easy to install the packages the project depends on if it is moved to another computer.

`drake` (OpenSci (2020)) analyzes workflows, skips steps where results are up to date, utilizes optimized computing to complete the rest of the steps, and provides evidence that results match the underlying code and data.

Lastly, the `reproducible` (McIntire & Chubaty (2020)) package focuses on the concept of caching: saving information so that projects can be run faster each time they are re-completed from the start.

There have also been several `Continuous integration` tools developed outside of R which can be used by those coding in almost any language. These provide more general approaches to automated checking, which can enhance reproducibility with minimal code.

For example, `wercker`—a command line tool that leverages Docker—enables users to test whether their projects will successfully compile when run on a variety of operating systems without access to the user's local hard drive (Oracle Corporation (2019)).

`GitHub Actions` is integrated into GitHub and can be configured to do similar checks on projects hosted in repositories.

`Travis CI` and `Circle CI` are popular continuous integration tools that can also be used to check `R` code.
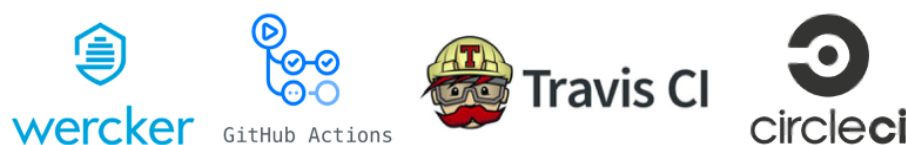
Chapter 2

`fertile`: My Contribution To Addressing Reproducibility

2.1 Understanding The Gaps In Existing Reproducibility Solutions

Although the current state of reproducibility in academia is quite poor, it is not an impossible challenge to overcome. The relative simplicity of addressing reproducibility, particularly when compared with replicability, makes it an ideal candidate for solution-building. Although significant progress on addressing reproducibility on a widespread scale is a long-term challenge, impactful forward progress–if on a smaller scale–can be achieved in the short-term. As we have seen, software developers, data scientists, and educators around the world have realized this potential, taking steps to help address the current crisis of reproducibility. Journals have put in place guidelines for authors, statisticians have developed `R` packages that help structure projects in a reproducible format, and educators have begun integrate reproducibility exercises into their courses.

However, many of these attempts to address reproducibility have significant drawbacks associated with them. We have already explored the issues with journal policies, both for authors and reviewers, in-depth. In this section, we will consider the education and software solutions and their associated challenges.

2.1.1 In Education

The two primary concerns about the integration of reproducibility in data science curricula revolve around time and difficulty.

As noted previously, the primary mode of teaching reproducibility is through the assignment of replication studies where students must take an existing study and go through the process of reproducing it themselves, including contacting the author for all necessary materials, rerunning code and analysis, and problem-solving when issues almost certainly come up.

In addition to the time required for the professor to collect all of the studies that

students will be working on, the inclusion of such an assignment places a significant burden on educators by taking up time where they could be teaching other important material. Replication studies, if done correctly, can take weeks for students to successfully complete. The choice to give such assignments is therefore associated with a significant opportunity cost which many professors are unwilling to take.

Additionally, both replication studies assigned in class and replication workshops outside of normal coursework require a working knowledge of how to successfully complete and understand research. This makes them inaccessible to individuals who are still in their undergraduate career and may not yet have had an opportunity to conduct research or those who are studying in non-research-focused technical programs.

In order to reach the widest variety of students possible, it is necessary to develop a new method of teaching reproducibility that is neither time consuming nor dependent on a prior understanding of the research process.

2.1.2　In Software

Previously, we considered several different types of software solutions: packages designed for users of `R` and continuous integration programs that can be used alongside a variety of coding languages. Although these solutions have their advantages, they also have significant drawbacks in terms of their ability to address reproducibility on a widespread scale.

Many of the packages designed for `R` are incredibly narrow in scope, with each effectively addressing a small component of reproducibility: file structure, modularization of code, version control, etc. They often succeed in their area of focus, but at the cost of accessibility to a wider audience. Their functions are often quite complex to use, and many steps must be completed to achieve the required reproducibility goal. This cumbersome nature means that most reproducibility packages currently available are not easily accessible to users with minimal `R` experience, nor particularly useful to those looking for quick and easy reproducibility checks. The significant learning curve associated with them can also detract potential users who may be interested in reproducibility but not willing to dedicate an extensive amount of time to understanding the intricacies of software operation.

Due to their generalized design, Continuous Integration tools do not face the same issues with narrowness or complexity that `R` packages struggle with. However, this generalizability provides its own additional challenge. Since Continuous Integration tools are designed to be accessible to a wide variety of users with different coding preferences, they are not particularly customizable and lack the ability to address features specific to certain programming languages.

Neither of these different software solutions appear to adequately address the challenge of reproducibility. In order to be the most effective, a piece of software must instead:

1) Be simple, with a small library of functions/tools that are straightforward to use.
2) Be accessible to a variety of users, with a relatively small learning curvey.

3) Be able to address a wide variety of aspects of reproducibility, rather than just one or two key issues.
4) Have features specific to a particular coding language that can address that language's unique challenges.
5) Be customizable, allowing users to choose for themselves which aspects of reproducibility they want to focus on.

2.2 `fertile`, An R Package Creating Optimal Conditions For Reproducibility

What if it were possible to address the existing issues with both educational and software reproducibility solutions simultaneously?

That is where my work comes in. In an attempt to produce meaningful change in the field of reproducibility, I have been developing `fertile`, a software package designed for `R` which helps users create optimal conditions for achieving reproducibility in their projects.

`fertile` attempts to address the gaps in existing reproducbility solutions by combining software and education in one product. The package provides a set of simple, easy-to-learn tools that, rather than focus intensely on a specific area like other software programs, provide some information about a wide variety of aspects influencing reproducibility. It is also designed to be incredibly flexible, offering benefits to users at any stage in the data analysis workflow and providing users with the option to select which aspects of reproducibility they want to focus on.

`fertile` also contains several`R`-specific features, which address certain aspects of reproducibility that can be missed by external project development tools. It is designed primarily to be used on data analyses organized as `R` Projects (i.e. directories containing an `.Rproj` file) and contains several associated features to ensure that the project structure meets the standards discussed in the `R` community.

In addition, `fertile` is designed to be educational, teaching its users about the components of reproducibility and how to achieve them in their work. The package provides users with detailed reports on the aspects of reproducibility where their projects fell short, identifying the root causes and, in many cases, providing a recommended solution.

`fertile` is structured in such a way as to be understandable and operable to individuals of any skill level, from students in their first undergraduate data science course to experienced PhD statisticians. The majority of its tools can be accessed in only a handful of functions with minimal required arguments. This simplicity makes the process of achieving and learning about reproducibility accessible to a wide audience in a way that complex software programs or graduate courses requiring an advanced knowledge of research methods do not.

Reproducibility is significantly easier to achieve when all of the tools necessary to do so are located in one place. `fertile` provides this optimal all-inclusive structure, addressing all six of the primary components of reproducibility discussed in Chapter

1. We will consider `fertile`'s treatment of each of these components in turn:

2.2.1   Component 1: Accessibility of Project Files

- Data (raw and/or processed)
- Metadata
- Code
- Related Software

2.2.2   Temporary Directories

The `sandbox()` function allows the user to make a copy of their project in a temporary directory. This can be useful for ensuring that projects run properly when access to the local file system is removed.

```
-- Attaching packages --------------------------------------------------- tidyverse 1.3.0 --

v ggplot2 3.3.2     v purrr   0.3.4
v tibble  3.0.3     v dplyr   1.0.2
v tidyr   1.1.2     v stringr 1.4.0
v readr   1.3.1     v forcats 0.5.0

-- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()


Attaching package: 'fertile'

The following object is masked from 'package:dplyr':

    tbl

The following objects are masked from 'package:readr':

    read_csv, read_csv2, read_delim, read_file, read_file_raw,
    read_fwf, read_lines, read_lines_raw, read_log, read_table,
    read_table2, read_tsv, write_csv

The following object is masked from 'package:ggplot2':

    ggsave

The following object is masked from 'package:stats':

    read.ftable

The following objects are masked from 'package:utils':

    read.csv, read.csv2, read.delim, read.delim2, read.DIF,
    read.fortran, read.fwf, read.table, write.csv

The following objects are masked from 'package:base':

    library, load, read.dcf, require, save, set.seed, setwd, source
```

```
fs::dir_ls('project_miceps') %>% head(3)
```

```
temp_dir <- sandbox('project_miceps')
temp_dir
fs::dir_ls(temp_dir) %>% head(3)
```

What we *have*:

- proj_report() gives the files present in the directory
- Every time a project is rendered, a session info file is created
- proj_pkg_script() for dependencies
- sandbox

What we *need*:

- can we report a checklist of the types of files we have available?
- it should check for the existence of a code file and data file and also readme

2.2.3   Component 2: Organized Project Structure

`fertile` provides a wide variety of features for managing the file system of a project. Nine of the package's fifteen primary reproducibility checks relate to file structure.

Two of these are focused on the `R project` aspect of the file system. `has_proj_root()` ensures that there is a single `.Rproj` file indicating a clear root directory for the project, while `has_no_nested_proj_root()` ensures that there are no sub-projects within. The recognition of a clear root directory is necessary to allow for file structure analysis and project restructuring as it provides a baseline directory to define relative file paths from.

Six of the major checks, whose names begin with `has_tidy_` focus on file clutter, addressing one of the components of a clean file structure. They check to make sure that no audio/video, image, source, raw data, .rda, or .R files are found in the root directory of the project.

The last check that focuses on the file system is `has_only_used_files()`. One of the more complicated checks in `fertile`, this function checks to make sure that there are no extraneous files serving no purpose (that are not "used") included alongside the analysis. For this function to work properly, a clear definition of which files are considered as "used" is needed. In `fertile`, that definition includes the following:

- `.R`,`.Rmd`, and `.Rproj` files
- README files
- Data/text/image files whose file path is referenced, either as an input or output, in any of the `.R` or `.Rmd` files
- Outputs created by knitting any `.Rmd` files (`.html`, `.pdf`, or `.docx` files with the same name as the `.Rmd`)
- Files created by `fertile`. In an effort to capture information about dependencies, `fertile` creates a text file capturing the session information when a

project is run and provides an option to generate an `.R` script that can install all of the packages referenced in a project's code. These files are considered "used."

```r
library(fertile)
```

```r
# List the files in the directory
fs::dir_ls("project_miceps")
```

```
project_miceps/Blot_data_updated.csv    project_miceps/CS_data_redone.csv
project_miceps/Estrogen_Receptors.docx  project_miceps/README.md
project_miceps/analysis.Rmd             project_miceps/analysis.html
project_miceps/citrate_v_time.png       project_miceps/mice.csv
project_miceps/miceps.Rproj             project_miceps/proteins_v_time.png
project_miceps/software-versions.txt
```

```r
# Check to see which are "used"
has_only_used_files("project_miceps")
```

```
Joining, by = "path_abs"

* Checking to see if all files in directory are used in code

  Problem: You have files in your project directory which are not being used.

  Solution: Use or delete files.

  See for help: ?fs::file_delete

# A tibble: 2 x 1
  path_abs
  <chr>
1 /Users/audreybertin/Documents/thesis/index/project_miceps/Estrogen_Receptors.~
2 /Users/audreybertin/Documents/thesis/index/project_miceps/mice.csv
```

There are also several functions that summarize and reshape the entire project structure. `proj_analyze_files()` lists all of the files in the project, along with their size, type, and relative path within the directory.

```r
proj_analyze_files("project_miceps")
```

```
# A tibble: 11 x 5
   file                  size ext   mime                  path_rel
   <fs::path>        <fs::by> <chr> <chr>                 <fs::path>
 1 project_miceps/Blot~  14.43K csv   text/csv              Blot_data_upd~
```

```
 2 project_miceps/CS_d~      7.39K csv    text/csv                     CS_data_redon~
 3 project_miceps/Estr~     10.97K docx   application/vnd.openxmlfo~   Estrogen_Rece~
 4 project_miceps/READ~        39 md      text/markdown               README.md
 5 project_miceps/anal~      5.1K Rmd     text/x-markdown             analysis.Rmd
 6 project_miceps/anal~     1.41M html    text/html                   analysis.html
 7 project_miceps/citr~    187.52K png    image/png                   citrate_v_tim~
 8 project_miceps/mice~     14.33K csv    text/csv                    mice.csv
 9 project_miceps/mice~       204 Rproj   text/rstudio                miceps.Rproj
10 project_miceps/prot~    380.28K png    image/png                   proteins_v_ti~
11 project_miceps/soft~      1.82K txt    text/plain                  software-vers~
```

`proj_suggest_moves()` provides suggestions for how the files should be reorganized, if at all.

```
files <- proj_analyze_files("project_miceps")

proj_suggest_moves(files)
```

```
# A tibble: 9 x 8
  file            size   ext   mime     path_rel    dir_rel    path_new    cmd
  <fs::path>   <fs::by> <chr> <chr>     <fs::path> <fs::path> <fs::path>  <chr>
1 project_mic~   14.43K csv   text/csv Blot_data~ data-raw    project_mi~ file_m~
2 project_mic~    7.39K csv   text/csv CS_data_r~ data-raw    project_mi~ file_m~
3 project_mic~   10.97K docx  applica~ Estrogen_~ inst/other  project_mi~ file_m~
4 project_mic~    5.1K Rmd    text/x-~ analysis.~ vignettes   project_mi~ file_m~
5 project_mic~    1.41M html  text/ht~ analysis.~ inst/text   project_mi~ file_m~
6 project_mic~  187.52K png   image/p~ citrate_v~ inst/image  project_mi~ file_m~
7 project_mic~   14.33K csv   text/csv mice.csv   data-raw    project_mi~ file_m~
8 project_mic~  380.28K png   image/p~ proteins_~ inst/image  project_mi~ file_m~
9 project_mic~    1.82K txt   text/pl~ software-~ inst/text   project_mi~ file_m~
```

These suggestions are based on achieving the optimal file structure design for reproducibility, argued by Marwick et al. (2018) to be that of an `R` package (R-Core-Team (2020), Wickham (2015)), with an `R` folder, as well as `data`, `data-raw`, `inst`, and `vignettes`. Such a structure keeps all of the files separated and in clearly labeled directories where they are easy to find. Additionally, the extension of the `R` package structure to data analysis projects promotes standardization of file structure within the R community. `proj_move_files()` executes these suggestions, building a package structure and sorting files into the correct folders by type.

Combined together, all of these functions make it simple to ensure that projects fall under a standard, simple, and reproducible structure with minimal clutter.

2.2.4   Component 3: Documentation

- Files are clearly named, preferably in a way where the order in which they should be run is clear.

- A README is present.
- Code contains comments.
- Software dependencies are noted.

High quality documentation, including the presence of a README file, comments explaining the code, a list of software packages an analysis is dependent on, and a method in which to understand which order to run code files in, is incredibly important to reproducibility. Without written guidance, individuals looking to reproduce results may not understand how to take all of the project components and combine them in the right way to produce the desired outcome. Additionally, code used without the proper dependencies and software versions, even if it is perfectly functional and correctly written, will often result in errors when compiled.

`fertile` contains a variety of functions to ensure that projects are well documented.

The `has_readme()` function ensures that a README file is included alongside a project.

```
has_readme("project_miceps")
```

v Checking for README file(s) at root level

`has_clear_build_chain()` looks to see if the project includes a `makefile`/`drakefile` (make-style file generated by the `Drake` package in `R`) or, if multiple code files are included, whether they are numbered. Any of these conditions being true would indicate that the order in which to run different pieces of code is clear. `project_miceps`, since it only contains one code file, passes this check easily:

```
has_clear_build_chain("project_miceps")
```

v Checking for clear build chain

There are also several features focused on software dependencies. Every time the code contained within a project is rendered/run by `fertile`, the package generates a file capturing the `sessionInfo()` just after the code was run. This file contains information about the R version in which the code was run, the list of packages that were loaded, and their specific versions.

The dependency information from `project_miceps` looks as follows:

```
session_info <- "project_miceps/software-versions.txt"
cat(readChar(session_info, 1e5))
```

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS Catalina 10.15.5
```

```
Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] stargazer_5.2.2  skimr_2.1.2       purrr_0.3.4       ggplot2_3.3.2
 [5] readr_1.3.1      tidyr_1.1.2       dplyr_1.0.2       stringr_1.4.0
 [9] knitr_1.29.5     thesisdown_0.1.0 remotes_2.2.0

loaded via a namespace (and not attached):
 [1] tidyselect_1.1.0   xfun_0.17          repr_1.1.0         splines_4.0.2
 [5] lattice_0.20-41    colorspace_1.4-1   vctrs_0.3.4        generics_0.0.2
 [9] usethis_1.6.1      htmltools_0.5.0    base64enc_0.1-3    yaml_2.2.1
[13] mgcv_1.8-31        utf8_1.1.4         rlang_0.4.7        pillar_1.4.6
[17] glue_1.4.2         withr_2.2.0        lifecycle_0.2.0    munsell_0.5.0
[21] gtable_0.3.0       evaluate_0.14      labeling_0.3       fansi_0.4.1
[25] highr_0.8          broom_0.7.0        Rcpp_1.0.5         backports_1.1.9
[29] scales_1.1.1       jsonlite_1.7.1     mime_0.9           farver_2.0.3
[33] fs_1.5.0           hms_0.5.3          digest_0.6.25      stringi_1.5.3
[37] bookdown_0.20.3    rprojroot_1.3-2    grid_4.0.2         fertile_0.0.0.9027
[41] here_0.1           cli_2.0.2          tools_4.0.2        magrittr_1.5
[45] tibble_3.0.3       crayon_1.3.4       pkgconfig_2.0.3    ellipsis_0.3.1
[49] Matrix_1.2-18      assertthat_0.2.1   rmarkdown_2.3      rstudioapi_0.11
[53] R6_2.4.1           nlme_3.1-148       compiler_4.0.2
```

The `proj_pkg_script()` makes it simple and fast to download these dependencies. When run on an R project directory, the function creates a `.R` script file that contains the code needed to install all of the packages referenced in the project, differentiating between packages located on CRAN and those located on GitHub.

```
install_script <- proj_pkg_script("project_miceps")
cat(readChar(install_script, 1e5))
```

```
# Run this script to install the required packages for this
R project.
# Packages hosted on CRAN...
install.packages(c( 'broom', 'dplyr', 'fs', 'ggplot2',
'purrr', 'readr', 'rmarkdown', 'skimr', 'stargazer',
'tidyr' ))
# Packages hosted on GitHub...
```

What we *need*:

- possibly a makefile generator? or improvements to order checking?

2.2.5   Component 4: File Paths

`fertile` contains a variety of features designed to address issues with file paths. Several of these features happen proactively, warning users that they are entering a non-reproducible file path as it happens, while others can be accessed after a project has already been written.

Proactively, `fertile` catches when an absolute path, or one leading outside of the project directory, is referenced in an input or output function and outputs a warning message. This interactive behavior is one of the educational features of the package, as it immediately provides an informative message explaining what went wrong, giving the programmer an opportunity to learn and course-correct from there.

```
library(fertile)
file.exists("~/Desktop/my_data.csv")
```

```
[1] TRUE
```

```
read.csv("~/Desktop/my_data.csv")
```

```
Error: Detected absolute paths
```

```
read.csv("../../../Desktop/my_data.csv")
```

```
Error: Detected paths that lead outside the project directory
```

`fertile` is particularly strict with the function `setwd()`, as its use is almost certain to break reproducibility. Calls to `setwd()` are blocked from executing, and users must conduct a manual override, via the `danger()` function, to change their working directory.

```
getwd()
```

```
[1] "/Users/audreybertin/Documents/thesis/index"
```

```
setwd("~/Desktop")
```

```
Error: setwd() is likely to break reproducibility. Use here::here() instead.
```

```
getwd()
```

```
[1] "/Users/audreybertin/Documents/thesis/index"
```

`fertile` also contains functionality that analyzes code for path information after it is written. For example, `proj_analyze_paths` goes through all of the paths used in a project's `.R` and `.Rmd` files and produces a report detailing which ones failed to meet reproducibility standards, explaining the problem, and providing a solution.

```
proj_analyze_paths('project_miceps')
```

```
-- Rendering R scripts... ---------------------------------------------- fertile 0.0.0.9027 --

Warning: 'data_frame()' is deprecated as of tibble 1.1.0.
Please use 'tibble()' instead.
This warning is displayed once every 8 hours.
Call 'lifecycle::last_warnings()' to see where this warning was generated.

-- Generating reproducibility report... --------------------------------- fertile 0.0.0.9027 --

Checking for absolute paths...

Checking for paths outside project directory...

New names:
* path -> path...1
* path -> path...5

# A tibble: 2 x 6
  path...1        path_abs        func   path...5        problem      solution
  <chr>           <chr>           <chr>  <chr>           <chr>        <chr>
1 /Users/audreyb~ /Users/audreyb~ readr~ /Users/audrey~  Absolute p~  Use a relat~
2 /Users/audreyb~ /Users/audreyb~ readr~ /Users/audrey~  Path is no~  Move the fi~
```

Paths can also be checked individually, or in groups, using `check_path()`.

```
# Path inside the directory
check_path("project_miceps")
```

```
# A tibble: 0 x 3
# ... with 3 variables: path <chr>, problem <chr>, solution <chr>
```

```
# Absolute path (current working directory)
check_path(getwd())
```

```
Error: Detected absolute paths
```

```
# Path outside the directory
check_path("../README.md")
```

```
Error: Detected paths that lead outside the project directory
```

These functions, together, cover all of the reproducibility subcomponents involving file paths. Users are warned not to use absolute paths or paths pointing outside of the directory, both while they are coding and after the fact, and provided with a recommended solution to the problem if they do so. Additionally, users are prevented from using functions that will certainly break reproducibility and provided with safer alternatives.

2.2.6   Component 5: Randomness

The component of randomness is addressed by `fertile` in a reproducibility check function titled `has_no_randomness()`, which does the following: First, it checks code files to determine whether they have employed random number generation. This is confirmed by recording the random number generation seed prior to running the files, rendering the files, and then checking to see whether the seed has changed. If the seed has changed, that indicates that random number generation has occurred. Then, it determines whether whether any randomness present was semi-random (i.e. controlled and repeatable) or truly random, by checking whether the `set.seed()` function was called. If the randomness was semi-random, it is considered permissible and reproducible. Truly random code is flagged as dangerous.

Let us consider an example from `project_miceps`. The primary code file, `analysis.Rmd` contains the following code. In this code, we see that a data set is read in via `read_csv` and reformatted. Then, the `sample_frac()` function takes a random sample of 50% of the rows of the data.

```
mice <- read_csv("Blot_data_updated.csv") %>%
  transmute(time = 'Time point',
            sex = Sex,
            id = 'Mouse #',
            erb = 'ERB Normalized Final',
            era = 'ERA Normalized Final',
            gper = 'GPER Normalized Final') %>%
  mutate(day = case_when(
    time == 0 ~ -1,
    time == 1 ~ 0,
    time == 2 ~ 1,
    time == 3 ~ 3,
    time == 4 ~ 5,
    time == 5 ~ 7
  ),
  type = ifelse(time == 0, "control", "treatment")
  )
mice_tidy <- mice %>%
  select(-time) %>%
  gather(key = "protein", value = "amount", -day, -sex, -id, -type) %>%
  mutate(protein_label = factor(protein,
    labels = c("paste(ER, alpha)", "paste(ER, beta)", "GPER")))
```

```
sample_frac(mice, 0.5)
```

However, there is no seed set in the file. As a result, `has_no_randomness()` returns as a failure for the project. Had a seed been set, however, the project would have passed.

```
has_no_randomness('project_miceps')
```

```
* Checking for no randomness

  Problem: Your code uses randomness

  Solution: Set a seed using `set.seed()` to ensure reproducibility.

  See for help: ?set.seed

# A tibble: 1 x 2
  culprit expr
  <chr>   <glue>
1 ?       Example: set.seed(1)
```

`fertile`'s randomness-centered feature helps analysts know that their use of randomness is controlled, ensuring that functions involving random number generation will always produce the same output each time they are run.

2.2.7   Component 6: Readability and Style

Though not an absolutely necessary component of reproducibility, code style can have a significant impact on how easy it is to follow the steps being taken in an analysis. The use of consistent and highly-readable style in code greatly simplifies and speeds up the process of understanding a data analysis project and repeating the steps included within.

  `fertile` addresses code style via the function `has_no_lint()`. This function checks code for compatibility with RStudio Chief Scientist Hadley Wikham's "tidy" code format, which promotes the following best practices:

- Line length should not exceed 80 characters
- There should not be trailing whitespace
- All infix operators (`=`, `+`, `-`, `<-`, etc) should have spaces on either side
- All commas should be followed by spaces
- There is no code left in the file that is commented out
- `<-` should be used instead of `=` to assign variables
- Opening curly braces should never go on their own line and should always be followed by a new line
- There should always be a space between right parenthesis and an open curly brace
- Closing curly braces should always be on their own line, unless followed by an `else` statement

Any issues with incompatibility that are found by `has_no_lint()` appear in a window by the RStudio console. This window lists the style inconsistencies found in each code document, showing both the description of the error and the line number where it

occurred for each issue. This window is interactive; double-clicking on an error brings the user to exact location in the code where it occurred.

Let us consider an example. `project_miceps` contains an `RMarkdown` file titled `analysis.Rmd`, which contains some code involved with a data analysis. Some of the code is in tidy style but not all of it is.

For example, line 71, part of a `ggplot` call, contains the following code:

```
# Line 71 of 'project_miceps/analysis.Rmd'

geom_hline(aes(yintercept = estimate + 1.96*std.error, color = sex), linetype = 3) +
```

For this line, `has_no_lint()` finds the following inconsistencies with tidy style:

Line 71 Lines should not be more than 80 characters.

Line 71 Put spaces around all infix operators.

Line 71 Trailing whitespace is superfluous.

The first comes up because the true line length, including spaces, is 84 characters. The second comes up because the `*` between `1.96` and `std.error` is not surrounded by spaces. The third is flagged because there is an empty space after the `+` at the end of the line.

An even less tidy piece of code is that found on line 189, written as follows:

```
# Line 189 of 'project_miceps/analysis.Rmd'

if (length(mice) > 1){ holder_var = 1 }
```

For this line, `has_no_lint()` finds the following errors:

Line 189 There should be a space between right parenthesis and an opening curly brace.

Line 189 Opening curly braces should never go on their own line and should always be followed by a new line.

Line 189 Use <-, not =, for assignment.

Line 189 Closing curly braces should always be on their own line, unless it's followed by an else.

The first issue flagged is the fact that `){`, in the middle of the line, has no space in the middle. The second issue is flagged because the code to execute when the `if` statement is true (`holder_var = 1`) is on the same line as the `if` statement, rather than on its own line. The third error occurs because `holder_var` is defined using an `=`. And the fourth occurs because the closing curly brace is on the same line as other code rather than by itself.

These informative error messages provided by `fertile` help teach users to use consistent, tidy style in their code. They do so while making the learning process incredibly simple, allowing users to click and see exactly which symbol or character caused the error. In addition to providing a detailed explanation of the was in which the provided code is not tidy, `fertile` also suggests a simple and fast solution for resolving the inconsistencies all at once, recommending that users apply the `usethis::use_tidy_style()` function to format their code automatically.

2.2.8 User Customizability

`fertile` does not force its users into a box, instead allowing for a great deal of user choice in terms of which aspects of reproducibility to focus on. Users can run reproducibility tests at three different scales:

1) Comprehensively, where all checks are run within a single function or two.
2) In groups, where functions focused on similar aspects of reproducibility are run together.
3) Individually, where only one reproducbility check is run at a time.

Users who want comprehensive behavior can access it through two primary functions, `proj_check()` and `proj_analyze()`.

The `proj_check()` function runs all fifteen reproducibility tests in the package, noting which ones passed, which ones failed, the reason for failure, a recommended solution, and a guide to where to look for help. These tests, many of which were described in detail previously, include: looking for a clear build chain, checking to make sure the root level of the project is clear of clutter, confirming that there are no files present that are not being directly used by or created by the code, and looking for uses of randomness that do not have a call to `set.seed()` present. A full list is provided below:

```
list_checks()
```

```
-- The available checks in 'fertile' are as follows: ---------------------- fertile 0.0.0.9027 --

 [1] "has_tidy_media"        "has_tidy_images"
 [3] "has_tidy_code"         "has_tidy_raw_data"
 [5] "has_tidy_data"         "has_tidy_scripts"
 [7] "has_readme"            "has_no_lint"
 [9] "has_proj_root"         "has_no_nested_proj_root"
[11] "has_only_used_files"   "has_clear_build_chain"
[13] "has_no_absolute_paths" "has_only_portable_paths"
[15] "has_no_randomness"
```

The `proj_analyze()` function creates a report documenting the structure of a data analysis project, combining four key functions from `fertile` into one complete unit:

- `proj_analyze_packages()`, which captures a list of all packages referenced in the source code and which files they were referenced in
- `proj_analyze_files()`, which captures a list of all of the files located in the directory along with their type and size
- `proj_suggest_moves()`, which provides suggestions for how to reorganize files to create a more reproducible file structure
- `proj_analyze_paths()`, which captures a list of the non-reproducible file paths referenced in source code

```
proj_analyze('project_miceps')
```

```
-- Analysis of reproducibility for project_miceps -------------------------- fertile 0.0.0.9027 --

--   Packages referenced in source code ------------------------------------ fertile 0.0.0.9027 --

# A tibble: 10 x 3
   package        N used_in
   <chr>      <int> <chr>
 1 broom          1 project_miceps/analysis.Rmd
 2 dplyr          1 project_miceps/analysis.Rmd
 3 fs             1 project_miceps/analysis.Rmd
 4 ggplot2        1 project_miceps/analysis.Rmd
 5 purrr          1 project_miceps/analysis.Rmd
 6 readr          1 project_miceps/analysis.Rmd
 7 rmarkdown      1 project_miceps/analysis.Rmd
 8 skimr          1 project_miceps/analysis.Rmd
 9 stargazer      1 project_miceps/analysis.Rmd
10 tidyr          1 project_miceps/analysis.Rmd

--   Files present in directory --------------------------------------------- fertile 0.0.0.9027 --

# A tibble: 12 x 4
   file              ext       size mime
   <fs::path>        <chr> <fs::byt> <chr>
 1 Estrogen_Receptor~ docx    10.97K application/vnd.openxmlformats-officedocu~
 2 citrate_v_time.png png    185.63K image/png
 3 proteins_v_time.p~ png    378.31K image/png
 4 Blot_data_updated~ csv     14.43K text/csv
 5 CS_data_redone.csv csv      7.39K text/csv
 6 mice.csv           csv     14.33K text/csv
 7 analysis.html      html     1.41M text/html
 8 README.md          md          39 text/markdown
 9 install_proj_pack~ R          251 text/plain
10 software-versions~ txt      2.27K text/plain
11 miceps.Rproj       Rproj      204 text/rstudio
12 analysis.Rmd       Rmd       5.1K text/x-markdown

--   Suggestions for moving files ------------------------------------------- fertile 0.0.0.9027 --

# A tibble: 10 x 3
   path_rel           dir_rel     cmd
   <fs::path>         <fs::path> <chr>
 1 Blot_data_updated~ data-raw   file_move('project_miceps/Blot_data_updated.cs~
 2 CS_data_redone.csv data-raw   file_move('project_miceps/CS_data_redone.csv',~
 3 Estrogen_Receptor~ inst/other file_move('project_miceps/Estrogen_Receptors.d~
 4 analysis.Rmd       vignettes  file_move('project_miceps/analysis.Rmd', fs::d~
 5 analysis.html      inst/text  file_move('project_miceps/analysis.html', fs::~
 6 citrate_v_time.png inst/image file_move('project_miceps/citrate_v_time.png',~
 7 install_proj_pack~ R          file_move('project_miceps/install_proj_package~
 8 mice.csv           data-raw   file_move('project_miceps/mice.csv', fs::dir_c~
 9 proteins_v_time.p~ inst/image file_move('project_miceps/proteins_v_time.png'~
10 software-versions~ inst/text  file_move('project_miceps/software-versions.tx~

--   Problematic paths logged ----------------------------------------------- fertile 0.0.0.9027 --
```

```
NULL
```

Together, `proj_check()` and `proj_analyze()` cover a significant portion of all six major components of reproducibility.

Users wanting to focus on groups of checks can do so using the `proj_check_some()` function. `proj_check_some()` leverages helper functions from `tidyselect` (Henry & Wickham (2020)) to allow users to call groups of similar checks together. `tidyselect` helpers tell `fertile` to call only the checks whose names meet certain conditions. The helper functions that can be passed to `proj_check_some()` are:

- `starts_with()`: Starts with a prefix.
- `ends_with()`: Ends with a suffix.
- `contains()`: Contains a literal string.
- `matches()`: Matches a regular expression.

For example, a variety of checks in `fertile` focus on making sure the project has a "tidy" structure–essentially that there are not files cluttered together all in one folder. Users interested in checking their tidyness can do so all at once using `proj_check_some()` as follows:

```
proj_check_some("project_miceps", contains("tidy"))
```

```
-- Running reproducibility checks ---------------------------------------- fertile 0.0.0.9027 --

v Checking for no *.R scripts at root level

v Checking for no *.rda files at root level

v Checking for no A/V files at root level

* Checking for no image files at root level

   Problem: Image files in root directory clutter project

   Solution: Move source files to img/ directory

   See for help: ?fs::file_move

# A tibble: 2 x 2
  culprit                  expr
  <fs::path>               <glue>
1 project_miceps/citrate_v_t~ fs::file_move('project_miceps/citrate_v_time.png'~
2 project_miceps/proteins_v_~ fs::file_move('project_miceps/proteins_v_time.png~

* Checking for no raw data files at root level

   Problem: Raw data files in root directory clutter project

   Solution: Move raw data files to data-raw/ directory

   See for help: ?fs::file_move

# A tibble: 3 x 2
  culprit                  expr
  <fs::path>               <glue>
1 project_miceps/Blot_data_u~ fs::file_move('project_miceps/Blot_data_updated.c~
2 project_miceps/CS_data_red~ fs::file_move('project_miceps/CS_data_redone.csv'~
3 project_miceps/mice.csv     fs::file_move('project_miceps/mice.csv', here::he~
```

```
v Checking for no source files at root level

-- Summary of fertile checks ---------------------------------------------- fertile 0.0.0.9027 --

v Reproducibility checks passed: 4

* Reproducibility checks to work on: 2

* Checking for no image files at root level

   Problem: Image files in root directory clutter project

   Solution: Move source files to img/ directory

   See for help: ?fs::file_move

# A tibble: 2 x 2
  culprit                   expr
  <fs::path>                <glue>
1 project_miceps/citrate_v_t~ fs::file_move('project_miceps/citrate_v_time.png'~
2 project_miceps/proteins_v_~ fs::file_move('project_miceps/proteins_v_time.png~

* Checking for no raw data files at root level

   Problem: Raw data files in root directory clutter project

   Solution: Move raw data files to data-raw/ directory

   See for help: ?fs::file_move

# A tibble: 3 x 2
  culprit                   expr
  <fs::path>                <glue>
1 project_miceps/Blot_data_u~ fs::file_move('project_miceps/Blot_data_updated.c~
2 project_miceps/CS_data_red~ fs::file_move('project_miceps/CS_data_redone.csv'~
3 project_miceps/mice.csv    fs::file_move('project_miceps/mice.csv', here::he~
```

Users might also attempt to call the two checks involving project roots together:

```
proj_check_some("project_miceps", ends_with("root"))
```

```
-- Running reproducibility checks --------------------------------------- fertile 0.0.

v Checking for nested .Rproj files within project

v Checking for single .Rproj file at root level

-- Summary of fertile checks ------------------------------------------- fertile 0.0.

v Reproducibility checks passed: 2
```

Or, perhaps, they might want to run all the functions beginning with "has_only":

```
proj_check_some("project_miceps", starts_with("has_only"))
```

-- Compiling... --- fertil

-- Rendering R scripts... --- fertil

Error in parse_block(g[-1], g[1], params.src, markdown_mode): Duplicate chunk label

-- Running reproducibility checks --------------------------------------- fertil

Joining, by = "path_abs"

v Checking for only portable paths

* Checking to see if all files in directory are used in code

   Problem: You have files in your project directory which are not being used.

   Solution: Use or delete files.

   See for help: ?fs::file_delete

```
# A tibble: 6 x 1
  path_abs
  <chr>
1 /Users/audreybertin/Documents/thesis/index/project_miceps/Blot_data_updated.c~
2 /Users/audreybertin/Documents/thesis/index/project_miceps/CS_data_redone.csv
3 /Users/audreybertin/Documents/thesis/index/project_miceps/Estrogen_Receptors.~
4 /Users/audreybertin/Documents/thesis/index/project_miceps/citrate_v_time.png
5 /Users/audreybertin/Documents/thesis/index/project_miceps/mice.csv
6 /Users/audreybertin/Documents/thesis/index/project_miceps/proteins_v_time.png
```

-- Summary of fertile checks --- fertil

v Reproducibility checks passed: 1

* Reproducibility checks to work on: 1

* Checking to see if all files in directory are used in code

   Problem: You have files in your project directory which are not being used.

   Solution: Use or delete files.

   See for help: ?fs::file_delete

```
# A tibble: 6 x 1
  path_abs
  <chr>
1 /Users/audreybertin/Documents/thesis/index/project_miceps/Blot_data_updated.c~
2 /Users/audreybertin/Documents/thesis/index/project_miceps/CS_data_redone.csv
3 /Users/audreybertin/Documents/thesis/index/project_miceps/Estrogen_Receptors.~
4 /Users/audreybertin/Documents/thesis/index/project_miceps/citrate_v_time.png
5 /Users/audreybertin/Documents/thesis/index/project_miceps/mice.csv
6 /Users/audreybertin/Documents/thesis/index/project_miceps/proteins_v_time.png
```

If users do not want to run functions in groups, and prefer to run them individually, that option is also provided to them. Every check that makes up `check()` and every subcomponent of `project_analyze()` can be run individually.

2.2.9   Educational Features

Simply noting and correcting issues with reproducibility is not enough to produce lasting change in the scientific community. Data analysts and software users must also be educated on why their choices were not reproducible so that they do not fall victim to those mistakes again in the future, but also so that they can share their knowledge and experience with others in the scientific community.

`fertile` prioritizes this idea of reproducibility education throughout many of its functionalities.

which records commands run in the console that have the potential to affect reproducibility, enabling users to look at their past history at any time. The package focuses mostly on package loading and file opening, noting which function was used, the path or package it referenced, and the timestamp at which that event happened. Users can access the log recording their commands at any time via the `log_report()` function:

```
log_report()
```

Reading from /Users/audreybertin/Documents/thesis/index/.fertile_log.csv

```
# A tibble: 11 x 4
   path            path_abs                        func       timestamp
   <chr>           <chr>                           <chr>      <dttm>
 1 package:remotes <NA>                            base::re~ 2020-09-19 17:35:37
 2 package:thesis~ <NA>                            base::re~ 2020-09-19 17:35:37
 3 package:thesis~ <NA>                            base::li~ 2020-09-19 17:35:37
 4 package:fertile <NA>                            base::li~ 2020-09-19 18:13:31
 5 package:stringr <NA>                            base::li~ 2020-09-19 18:13:46
 6 package:fertile <NA>                            base::li~ 2020-09-19 18:13:46
 7 ~/Desktop/my_d~ ~/Desktop/my_data.csv           utils::r~ 2020-09-19 18:13:46
 8 ../../../Deskt~ /Users/audreybertin/Desktop/my~ utils::r~ 2020-09-19 18:13:46
 9 package:purrr   <NA>                            base::li~ 2020-09-19 18:13:57
10 package:forcats <NA>                            base::li~ 2020-09-19 18:13:57
11 project_miceps~ /Users/audreybertin/Documents/~ readr::r~ 2020-09-19 18:13:57
```

The log, if not managed, can grow very long over time. For users who do not desire such functionality, `log_clear()` provides a way to erase the log and start over.

2.2.10 File Types

There are several functions that can be used to check the type of a file:

```
is_data_file(fs::path('project_miceps', "mice.csv"))
```

```
[1] TRUE
```

```
is_image_file(fs::path('project_miceps', "proteins_v_time.png"))
```

```
[1] TRUE
```

```
is_text_file(fs::path('project_miceps', "README.md"))
```

```
[1] TRUE
```

```
is_r_file(fs::path('project_miceps', "analysis.Rmd"))
```

```
[1] TRUE
```

2.3 How `fertile` Works

Much of the functionality in `fertile` is achieved by writing **shims link to wikipedia page here**. `fertile`'s shimmed functions intercept the user's commands and perform various logging and checking tasks before executing the desired function. Our process is:

1. Identify an `R` function that is likely to be involved in operations that may break reproducibility. Popular functions associated with only one package (e.g., `read_csv()` from `readr`) are ideal candidates.

2. Create a function in `fertile` with the same name that takes the same arguments (and always the dots ...).

3. Write this new function so that it:

a) captures any arguments,
b) logs the name of the function called,
c) performs any checks on these arguments, and
d) calls the original function with the original arguments. Except where warranted, the execution looks the same to the user as if they were calling the original function.

Most shims are quite simple and look something like what is shown below for `read_csv()`.

```
fertile::read_csv
```

```
function(file, ...) {
  if (interactive_log_on()) {
    log_push(file, "readr::read_csv")
    check_path_safe(file)
    readr::read_csv(file, ...)
  }
}
<bytecode: 0x7fc09ce7ead8>
<environment: namespace:fertile>
```

`fertile` shims many common functions, including those that read in a variety of data types, write data, and load packages. This works both proactively and retroactively, as the shimmed functions written in `fertile` are activated both when the user is coding interactively and when a file containing code is rendered.

In order to ensure that the `fertile` versions of functions ("shims") always supersede ("mask") their original namesakes when called, `fertile` uses its own shims of the `library` and `require` functions to manipulate the R search path so that it is always located in the first position. In the `fertile` version of `library()`, we detach `fertile` from the search path, load the requested package, and then re-attach `fertile`. This ensures that when a user executes a command, R will check `fertile` for a matching function before considering other packages. While it is possible that this shifty behavior could lead to unintended consequences, our goal is to catch a good deal of problems before they become problematic. Users can easily disable `fertile` by detaching it, or not loading it in the first place.

2.4 `fertile` in Practice: Experimental Results From Smith College Student Use

`fertile` is designed to: 1) be simple enough that users with minimal R experience can use the package without issue, 2) increase the reproducibility of work produced by its users, and 3) educate its users on why their work is or is not reproducible and provide guidance on how to address any problems.

To test `fertile`'s effectiveness, we began an initial randomized control trial of the package on an introductory undergraduate data science course at Smith College in Spring 2020 **ADD FOOTNOTE** (This study was approved by Smith College IRB, Protocol #19-032).

The experiment was structured as follows:

1.Students are given a form at the start of the semester asking whether they consent to participate in a study on data science education. In order to successfully consent, they must provide their system username, collected through the command `Sys.getenv("LOGNAME")`. To maintain privacy the results are then transformed into a hexadecimal string via the `md5()` hashing function.

2. These hexadecimal strings are then randomly assigned into equally sized groups,

one experimental group that receives the features of `fertile` and one group that receives a control.

3. The students are then asked to download a package called `sds192` (the course number and prefix), which was created for the purpose of this trial. It leverages an `.onAttach()` function to scan the `R` environment and collect the username of the user who is loading the package and run it through the same hashing algorithm as used previously. It then identifies whether that user belongs to the experimental or the control group. Depending on the group they are in, they receive a different version of the package.

4. The experimental group receives the basic `sds192` package, which consists of some data sets and `R` Markdown templates necessary for completing homework assignments and projects in the class, but also has `fertile` installed and loaded silently in the background. The package's proactive features are enabled, and therefore users will receive warning messages when they use absolute or non-portable paths or attempt to change their working directory. The control group receives only the basic `sds192` package, including its data sets and `R` Markdown templates. All students from both groups then use their version of the package throughout the semester on a variety of projects.

5. Both groups are given a short quiz on different components of reproducibility that are intended to be taught by `fertile` at both the beginning and end of the semester. Their scores are then compared to see whether one group learned more than the other group or whether their scores were essentially equivalent. Additionally, for every homework assignment submitted, the professor takes note of whether or not the project compiles successfully.

Based on the results, we hope to determine whether `fertile` was successful at achieving its intended goals. A lack of notable difference between the *experimental* and *control* groups in terms of the number of code-related questions asked throughout the semester would indicate that `fertile` achieved its goal of simplicity. A higher average for the *experimental* group in terms of the number of homework assignments that compiled successfully would indicate that `fertile` was successful in increasing reproducibility. A greater increase over the semester in the reproducibility quiz scores for students in the *experimental* group compared with the *control* group would indicate that `fertile` achieved its goal of educating users on reproducibility. Success according to these metrics would provide evidence showing `fertile`'s benefit as tool to help educators introduce reproducibility concepts in the classroom.

Chapter 3

Incorporating Reproducibility Tools Into The Greater Data Science Community

3.1 Potential Applications of `fertile`

3.1.1 In Journal Review

3.1.2 By Beginning Data Scientists

3.1.3 By Advanced Data Scientists

3.1.4 For Teaching Reproducibility

Nicole Janz – Brining the Gold Standard into the Classroom: Replication in University Teaching

3.2 Integration Of `fertile` And Other Reproducibility Tools in Data Science Education

Conclusion

`fertile` is an `R` package that lowers barriers to reproducible data analysis projects in `R`, providing a wide array of checks and suggestions addressing many different aspects of project reproducibility, including file organization, file path usage, documentation, and dependencies. `fertile` is meant to be educational, providing informative error messages that indicate why users' mistakes are problematic and sharing recommendations on how to fix them. The package is designed in this way so as to promote a greater understanding of reproducibility concepts in its users, with the goal of increasing the overall awareness and understanding of reproducibility in the `R` community.

The package has very low barriers to entry, making it accessible to users with various levels of background knowledge. Unlike many other `R` packages focused on reproducibility that are currently available, the features of `fertile` can be accessed almost effortlessly. Many of the retroactive features can be accessed in only two lines of code requiring minimal arguments and some of the proactive features can be accessed with no additional effort beyond loading the package. This, in combination with the fact that `fertile` does not focus on one specific area of reproducibility, instead covering (albeit in less detail) a wide variety of topics, means that `fertile` makes it easy for data analysts of all skill levels to quickly gain a better understanding of the reproducibility of the work.

In the moment, it often feels easiest to take a shortcut—to use an absolute path or change a working directory. However, when considering the long term path of a project, spending the extra time to improve reproducibility is worthwhile. `fertile`'s user-friendly features can help data analysts avoid these harmful shortcuts with minimal effort.

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readability and/or setup.

In the main Rmd file

```r
# This chunk ensures that the thesisdown package is
# installed and loaded. This thesisdown package includes
# the template files for the thesis.
if (!require(remotes)) {
  if (params$`Install needed packages for {thesisdown}`) {
    install.packages("remotes", repos = "https://cran.rstudio.com")
  } else {
    stop(
      paste('You need to run install.packages("remotes")',
            "first in the Console.')
    )
  }
}
```

Reading from /Users/audreybertin/Documents/thesis/index/.fertile_log.csv

```r
if (!require(thesisdown)) {
  if (params$`Install needed packages for {thesisdown}`) {
    remotes::install_github("ismayc/thesisdown")
  } else {
    stop(
      paste(
        "You need to run",
        'remotes::install_github("ismayc/thesisdown")',
        "first in the Console."
      )
    )
```

```
  }
}
```

Reading from /Users/audreybertin/Documents/thesis/index/.fertile_log.csv

```
library(thesisdown)
```

Reading from /Users/audreybertin/Documents/thesis/index/.fertile_log.csv

```
# Set how wide the R output will go
```

Appendix B

The Second Appendix, for Fun

References

American Economic Association. (2020). Data and code availability policy. Retrieved from `https://www.aeaweb.org/journals/data/data-code-policy`

American Journal of Political Science. (2016, May). Guidelines for preparing replication files. Retrieved from `https://ajps.org/wp-content/uploads/2018/05/ajps_replication-guidelines-2-1.pdf`

American Statistical Association. (2020). JASA acs reproducibility guide. Retrieved from `https://jasa-acs.github.io/repro-guide/pages/author-guidelines`

Baker, M. (2015). Over half of psychological studies fail reproducibility test. *Nature*. Retrieved from `https://www.nature.com/news/over-half-of-psychology-studies-fail-reproducibility-test-1.18248`

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*. Retrieved from `https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970`

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533.

Blischak, J., Carbonetto, P., & Stephens, M. (2019). Workflowr: A framework for reproducible and collaborative data science. Retrieved from `https://CRAN.R-project.org/package=workflowr`

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., Olds, J. L., & Dean, H. (2015). Report of the subcommittee on replicability in science advisory committee to the nsf sbe directorate.

Broman, K. (2019). Initial steps toward reproducible research: Organize your data and code. *Sitewide ATOM*. Retrieved from `https://kbroman.org/steps2rr/pages/organize.html`

Cambridge University Press. (2020). Experimental results - transparency and openness policy. Retrieved from `https://www.cambridge.org/core/journals/experimental-results/information/transparency-and-openness-policy`

Claerbout, J. F., & Karrenbach, M. (1992). Electronic documents give reproducible research a new meaning. In *SEG technical program expanded abstracts 1992* (pp.

601–604). Society of Exploration Geophysicists.

Cooper, N., Hsing, P.-Y., Croucher, M., Graham, L., James, T., Krystalli, A., & Michonneau, F. (2017). A guide to reproducible code in ecology and evolution. *British Ecological Society*. Retrieved from `https://www.britishecologicalsociety.org/wp-content/uploads/2017/12/guide-to-reproducible-code.pdf`

Eisner, D. A. (2018). Reproducibility of science: Fraud, impact factors and carelessness. *Journal of Molecular and Cellular Cardiology*, *114*, 364–368. `http://doi.org/https://doi.org/10.1016/j.yjmcc.2017.10.009`

Fidler, F., & Wilcox, J. (2018). Reproducibility of scientific results. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2018). `https://plato.stanford.edu/archives/win2018/entries/scientific-reproducibility/`; Metaphysics Research Lab, Stanford University.

FitzJohn, R., Ashton, R., Hill, A., Eden, M., Hinsley, W., Russell, E., & Thompson, J. (2020). Orderly: Lightweight reproducible reporting. Retrieved from `https://CRAN.R-project.org/package=orderly`

Gancarz, M. (2003). *Linux and the unix philosophy* (2nd ed.). Woburn, MA: Digital Press.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 1–6. `http://doi.org/10.1126/scitranslmed.aaf5027`

Gosselin, R.-D. (2020). Statistical analysis must improve to address the reproducibility crisis: The access to transparent statistics (acts) call to action. *BioEssays*, *42*(1), 1900189. `http://doi.org/10.1002/bies.201900189`

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., ... others. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal cognition. *Royal Society Open Science*, *5*(8), 180448. Retrieved from `https://royalsocietypublishing.org/doi/full/10.1098/rsos.180448`

Henry, L., & Wickham, H. (2020). Tidyselect: Select from a set of strings. Retrieved from `https://CRAN.R-project.org/package=tidyselect`

Hermans, F., & Aldewereld, M. (2017). Programming is writing is programming. In *Companion to the first international conference on the art, science and engineering of programming* (pp. 1–8).

Hrynaszkiewicz, I. (2020). Publishers' responsibilities in promoting data quality and reproducibility. *Handbook of Experimental Pharmacology*, *257*, 319–348. `http://doi.org/https://doi.org/10.1007/164_2019_290`

Jacoby, W. G., Lafferty-Hess, S., & Christian, T.-M. (2017). Should journals be responsible for reproducibility? Inside Higher Ed. Retrieved from

`https://www.insidehighered.com/blogs/rethinking-research/should-journals-be-responsible-reproducibility`

Journal of Computational and Graphical Statistics. (2020). Instructions for authors. Retrieved from `https://www.tandfonline.com/action/authorSubmission?show=instructions&journalCode=ucgs20`

Journal of Statistical Software. (2020). Instructions for authors. Retrieved from `https://www.jstatsoft.org/pages/view/authors#review-process.`

Kitzes, J., Turek, D., & Deniz, F. (2017). *The practice of reproducible research: Case studies and lessons from the data-intensive sciences.* Berkeley, CA: University of California Press. Retrieved from `https://www.practicereproducibleresearch.org`

Leopold, S. S. (2015). Editorial: Increased manuscript submissions prompt journals to make hard choices. *Clinical Orthopaedics and Related Research*, *473*(3), 753–755. `http://doi.org/10.1007/s11999-014-4129-1`

Martinez, C., Hollister, J., Marwick, B., Szöcs, E., Zeitlin, S., Kinoshita, B. P., . . . Meinke, B. (2018). Reproducibility in Science: A Guide to enhancing reproducibility in scientific results and writing. Retrieved from `http://ropensci.github.io/reproducibility-guide/`

Marwick, B. (2019). Rrtools: Creates a reproducible research compendium. Retrieved from `https://github.com/benmarwick/rrtools`

Marwick, B., Boettiger, C., & Mullen, L. (2018). Packaging data analytical work reproducibly using R (and friends). *The American Statistician*, *72*(1), 80–88. `http://doi.org/doi.org/10.1080/00031305.2017.1375986`

McArthur, S. L. (2019). Repeatability, reproducibility, and replicability: Tackling the 3R challenge in biointerface science and engineering. *Biointerphases*, *14*(2), 1–2. `http://doi.org/10.1116/1.5093621`

McIntire, E. J. B., & Chubaty, A. M. (2020). Reproducible: A set of tools that enhance reproducibility beyond package management. Retrieved from `https://CRAN.R-project.org/package=reproducible`

National Institutes of Health. (2014, June). Principles and guidelines for reporting preclinical research. Retrieved from `https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research`

OpenSci, R. (2020). Drake: A pipeline toolkit for reproducible computation at scale. Retrieved from `https://cran.r-project.org/package=drake`

Oracle Corporation. (2019). Wercker. Retrieved from `https://github.com/`

```
wercker/wercker
```

R Journal Editors. (2020). Instructions for authors. Retrieved from `https://journal.r-project.org/share/author-guide.pdf`

R-Core-Team. (2020). Writing r extensions. *R Foundation for Statistical Computing*. Retrieved from `http://cran.stat.unipd.it/doc/manuals/r-release/R-exts.pdf`

Ross, N., DeCicco, L., & Randhawa, N. (2018). Checkers: Automated checking of best practices for research compendia. Retrieved from `https://github.com/ropenscilabs/checkers/blob/master/DESCRIPTIONr`

Stodden, V., Seiler, J., & Ma, Z. (2018a). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584–2589. Retrieved from `https://www.pnas.org/content/115/11/2584`

Stodden, V., Seiler, J., & Ma, Z. (2018b). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584–2589. `http://doi.org/10.1073/pnas.1708290115`

The American Statistician. (2020). Instructions for authors. Retrieved from `https://www.tandfonline.com/action/authorSubmission?show=instructions&journalCode=utas20`

Ushey, K., & RStudio. (2020). Renv: Project environments. Retrieved from `https://cran.r-project.org/web/packages/renv/index.html`

Wallach, J. D., Boyack, K. W., & Ioannidis, J. P. A. (2018). Reproducible research practices, transparency, and open access data in the biomedical literature, 2015-2017. *PLOS Biology*, *16*(11), 1–20. `http://doi.org/10.1371/journal.pbio.2006930`

Wickham, H. (2015). *R packages* (1st ed.). Sebastopol, CA: O'Reilly Media, Inc.

Woolston, C. (2020). TOP factor rates journals on transparency, openness. Nature Index. Retrieved from `https://www.natureindex.com/news-blog/top-factor-rates-journals-on-transparency-openness`