

When is done *done*? Practice fundamentals of automated testing for collaborative data analysis development

How to go from no tests, to *whoa* tests!

sketch-draft by lead author, Charles T. Gray

Version of this manuscript

This manuscript will go through three phases:

1. A lead author sketch draft.
2. Implementation with the reproducibility team for The repliCATS Project & version listing any team members who wish to co-author.
3. Implementation of repliCATS procedures, explicitly, after we need not blind the algorithms for experimental integrity. Invitation for external co-authors to contribute and extend the content, review, and final submission. Possibly for useR2020.

Coding to doneness

Contrapunctus I is rarely learnt in one sitting, indeed, for most pianists, to even play the piece is the result of many months' diligent work, and still it will never feel *done*. For the student of this opening work of Bach's *The Art of Fugue*, the time at the keyboard spent to play through a three-minute piece of music can seem immeasurable. To play well, a diligent pianist must employ technique. But *how* a pianist is to achieve the desired technique is not clear (Chang 2009).

For a researcher performing an analysis using a computational tool such as R, it can be unclear when an analysis is *done*. And, while there are technical guides on *good enough* practice, it can feel overwhelming as to what to adopt and where. Particularly for an in-development algorithm, which is to say, an algorithm with scripts already started, and some results explored.

This manuscript picks up from where Riederer left off with *RMarkdown-driven development* and suggests a *coderegistering test-driven workflow* for coding to doneness. This workflow provides a roadmap to completion for a packaged analysis with `code::proofed` the manuscript (???), provided measures of confidence in the implementation of the algorithm.

Musicians and, similarly, athletes, do not see themselves as having mastered a skill, but active practitioners of a craft (Galway 1990). To be a flautist is to practice, and to be athlete is to train. For researcher developers, it can be hard to assess when an analysis is completed. In this manuscript, we consider the *practice* of test-driven analysis development as a means of *coding to doneness* for an algorithm, a workflow that facilitates.

Questionable research practices in scientific computing

Algorithms are coded by people who practice code, and significant problems emerge when algorithms are treated as fixed artifacts, rather than one of the tools utilised by those who *practice* code. In Australia, a heartbreakingly-ongoing example of this kind of problem with the income-reporting data analysis algorithm

that assesses if income welfare recipients have been overpaid entitlements. Crude averaging calculations have lead to ongoing incorrect debt notices issued, such as 20,000 people receiving “robodebt” notices for debts they did not owe (McIlroy 2017). Problems in data analysis have real impacts on real people’s lives. Perhaps, if the algorithm were considered a workflow practiced and monitored by a team of data scientists, rather than a static object, problems would not be persisting to this day (Karp 2019). Indeed, as noted in Wilson’s testing primers (in development) for RStudio (“RStudio Cloud,” n.d.),

Almost all (92%) of the catastrophic system failures are the result of incorrect handling of non-fatal errors explicitly signaled in software. In 58% of the catastrophic failures, the underlying faults could easily have been detected through simple testing of error handling code (Yuan et al. 2014).

We might view this as a computational instantiation of what Fraser *et al.* denote *questionable research practices* (QRPs). The QRPs Fraser *et al.* provide a taxonomy for refer to various practices in scientific methodology found in ecology and evolution, such as *p*-hacking, adjusting a model’s specification to achieve a desired *p*-value, or *cherry picking*, failing to report variables or results that do not fit a desired narrative. QRPs are, importantly, often not a result of scientific malpractice, but a question of discipline-specific conventions established in bygone eras not translating well to data-driven research (Fraser et al. 2018). In scientific computing, as the robodebt example illustrates, similar and overlapping errors may occur.

A consideration when providing recommendations of best practice is what we might reasonably expect of a researcher. Indeed, it is likely unrealistic to expect *best practices* in scientific computing (Wilson et al. 2014), perhaps we would be better off asking for *good enough* practices (Wilson et al. 2017) in scientific computing. For while researchers use computational algorithms in science every day, most of them are not trained in computational science. Even mathematicians and statisticians do a great deal of training at the blackboard, rather than the computer.

Riederer identifies several qrps in scientific computing, for example, hardcoded values that interfere with another’s ability to reproduce the computational results (Riederer 2019). As researcher software engineers, it behoves us to consider what are questionable research practices in software produced for data analyses. Version control and open code sharing via a platform such as GitHub, is one way to mitigate questionable research practices in scientific computing (Bryan 2017). There is also a growing literature on reproducible research compendia via packaged analyses (Marwick, Boettiger, and Mullen 2018; Wilson et al. 2017).

This manuscript contributes to this literature by focussing on workflows for using automated tests to move analysis a fully reproducible packaged research compendium. Furthermore, these workflow assist the developer communicate what they have coded to others and their future self, as a means of mitigating questionable research practices in scientific computing. Above all, by providing a workflow for the *practice* of code, the developer anxiety is reduced by having the parameters as clearly defined as they can be from the start.

Practice fundamentals

Chuang C. Chang’s *Fundamentals of Piano Practice* sets out to address a gap in piano pedagogy (Chang 2009). A similar gap exists in the implementation of *good enough* practices (Wilson et al. 2017), which is to say, what we might reasonably expect of a analyst, in reproducible computing for research data analyses. The objective is different from advanced pianism, but we characterise testing *practice*, as opposed to *technique*, analogous to how Chang delineates between *piano* practice and technique. Through attempting to identify the fundamentals of automated testing for collaborative data analysis development, this manuscript aims to articulate the gap in understanding automated testing implementation for analysts.

As Chang notes, whilst there a rich history of technical pedagogy, there is a dearth of guidance for pianists on *learning* the technique (Chang 2009). There are many canonical texts of pianistic technique pedagogy. Bach provides a pathway from small canons [todo: cite], to two-part inventions[*todo: cite*], three-part sinfonia, and the challenge of *The Well-Tempered Clavier* and *The Art of Fugue*. Bartok provides the *Mikrokosmos* [todo:], and Czerny’s *School of Velocity* [to do]. In each case, technical exercises of increasing difficulty are

provided. In piano pedagogy, a *technical* exercise isolates a particular aspect of pianistic technique [todo examples]. For example, [todo Czerny staccato]. Or, [todo voicing technique].

The dearth that Chang attempts to address is in pianistic *practice* habits that will lead to succesful adoption of these techniques. In science, we might call this *work flow* [todo cite].

.. practically every piano learning method consists of showing students what to practice, and what kinds of techniques (runs, arpeggios, legato, staccato, trills, etc.) are needed. There are few instructions on how to practice order to be able to play them, which is mostly left to the student and endless repetitions (Chang 2009).

Wilson et al. followed their work on *best practices* in scientific computing (Wilson et al. 2014), with a reflection on *good enough* practices (Wilson et al. 2017), in recognition that we must account for what we might reasonably request of practitioners of data analysis. In this manuscript, we consider one component of *good enough* practice in data analysis: *automated testing*.

Automated tests are a formalised way of implementing checks that inputs and outputs of algorithms are as expected (Wickham 2015). Informative messages are output that assist the developer in identifying where code is not performing as expected.

Collaboration via automated testing

At heart, automated tests are collaborative. This may be with others, but applies at least as much to a analyst approaching their own past work with which they have become unfamiliar, or have become anxious about some aspect of the work. Automated tests provide an efficient way of returning to and extending an analysis; anxiety is reduced by having defined outcomes to code explicitly for.

Reproducible research compendia provide a means by which analysts can share their work so that it may be extended by others, and automated tests provide `code::proof` (Gray 2019), a measure of confidence in the algorithm for others. Hayes expresses concern about using algorithms published without automated tests (Hayes 2019). However, only one quarter of the largest repository of R packages, The Comprehensive R Archive Network, have any automated tests (Hester 2016), highlighting that, despite testing identified as a ‘vital’ part of data analysis (Wickham 2015), automated testing is yet to be widely adopted.

A test-driven toolchain walkthrough for an in-development analysis

In this section we now consider the practicality of implementing tests through a *toolchain walkthrough*, an opinionated documetation of a scientific workflow , towards a measure of `code::proof`, confidence in the algorithm implemented (Gray 2019). In particular, a toolchain walkthrough is a reflection of *one* workflow, whilst others, indeed, better, workflows might exist. This is in contrast to a comprehensive review of tools. Instead, a toolchain walkthrough ruminates on considerations *after* a set of tools have been chosen.

Toolchain walkthroughs aim to identify obstacles and advantages in implementation of scientific workflows. By necessity, a toolchain walkthrough is language specific, but, ideally, observations emerge that are useful for any appropriate language employed for data analysis, such as Python. This toolchain walkthrough examines the *process*, analogous to piano *practice*, of implementing tests, as opposed to defining comprehensively the nature of *good enough* tests, analogous to guidance on pianistic technique.

Objective of this toolchain walkthrough

This toolchain walkthrough aims to provide guidance on implementing a test-driven workflow for an in-development analysis. Many analyses begin as scripts that develop [todo expan] (Riederer 2019). The central question of this manuscript is what constitutes a minimal level of testing for in-development analysis, where

code is still being written and features implemented. Automated tests assist in time-consuming debugging tasks (Wickham 2015), but also in providing information with a developer who is unfamiliar with the code.

This is a first effort in identifying the fundamentals of automated testing for the collaborative process of developing an analysis in R. Analogous to Riederer’s *RMarkdown-driven development* (Riederer 2019), which deconstructs the workflow of developing an analysis from .Rmd notebook-style reports to packaged analyses, we consider a set a computational tools that form a workflow to assist in the coherent development of automated tests for data analysis. This is an extension of the workflow suggestions provided in *R Packages*, with a specific focus on collaborative workflows in research.

Devops and assumed expertise

This toolchain walkthrough assumes a knowledge of R advanced enough to be using the language to be answering scientific research claims.

- tools used: testthat, neet, covr
- GitHub

Get an overview with `covr::`

With a packaged analysis, it’s easy to get a sense of existing automated testing coverage, if any tests have been written. The analysis functions provided by `varameta::` were put on hold while significant discussion was written.

```
library(covr)
package_coverage("~/Documents/repos/varameta/")
#> varameta Coverage: 70.71%
#> R/dist_name.R: 0.00%
#> R/g_cauchy.R: 44.44%
#> R/g_norm.R: 71.43%
#> R/hozo_se.R: 92.31%
#> R/bland_mean.R: 100.00%
#> R/bland_se.R: 100.00%
#> R/effect_se.R: 100.00%
#> R/g_exp.R: 100.00%
#> R/g_lnorm.R: 100.00%
#> R/hozo_mean.R: 100.00%
#> R/wan_mean_C1.R: 100.00%
#> R/wan_mean_C2.R: 100.00%
#> R/wan_mean_C3.R: 100.00%
#> R/wan_se_C1.R: 100.00%
#> R/wan_se_C2.R: 100.00%
#> R/wan_se_C3.R: 100.00%
```

Created on 2020-02-01 by the reprex package (v0.3.0)

The `covr::` function `package_coverage` provides the percentage of the lines of code run in tests.

Code with intent

When we think of an algorithm, it’s easy to feel overwhelmed by the complexity. Here are the relationships between inputs, and just some of the estimators, and outputs in `varameta`, presented as a randomised *graph*, a visual representation of a set of nodes, V , and the edges, $V \times V$, between them.

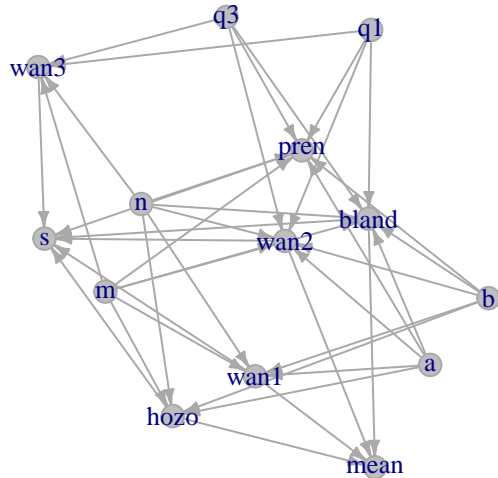
```

library(igraph)

confusing_relationships <- make_graph(~
a-->hozo, b-->hozo, m-->hozo, n-->hozo, hozo-->s, hozo-->mean,
a-->wan1, b-->wan1, m-->wan1, n-->wan1, wan1-->s, wan1-->mean,
a-->pren, b-->pren, m-->pren, n-->pren, pren-->s,
q1-->wan3, q3-->wan3, m-->wan3, n-->wan3, wan3-->s, wan3-->mean,
q1-->pren, q3-->pren, m-->pren, n-->pren, pren-->s,
a-->wan2, b-->wan2, q1-->wan2, q3-->wan2, m-->wan2, n-->wan2, wan2-->s, wan2-->mean,
a-->bland, b-->bland, q1-->bland, q3-->bland, m-->bland, n-->bland, bland-->s, bland-->mean
)

confusing_relationships %>% plot(
  vertex.color = "grey",
  vertex.frame.color = "darkgrey",
  vertex.size = 10,
  vertex.label.cex = 0.8,
  edge.arrow.size = 0.4
)

```



The Center for Open Science recommend *preregistering* an experiment, stating what the hypothesis is and how the analyst intends to assess the hypothesis as one safeguard against inadvertent questionable research practices. Analogously, we might think of preregistering our code as providing some `code::proof` of the implementation of our algorithm. In this manuscript, we will think of *coderegistration*, stating what the intention of an algorithm is, as

In the 2000s, Francesca Cann taught concert-pianists in training to *practice with intent*, to do what one has intended to do.

Coderegistering an algorithm

For this coderegistration, we To coderegister an algorithm:

On a piece of paper.

1. Draw a diagram of the inputs and outputs of the algorithm.
2. Draw where this algorithm fits in the pipeline of the package, if appropriate.

In an issue on GitHub:

1. Describe the algorithm's intended purpose.
2. Describe the input parameters and how they will be tested.
3. Describe the output parameters and how they will be tested.

Update the coderegistration as needed.

Sketch-diagram

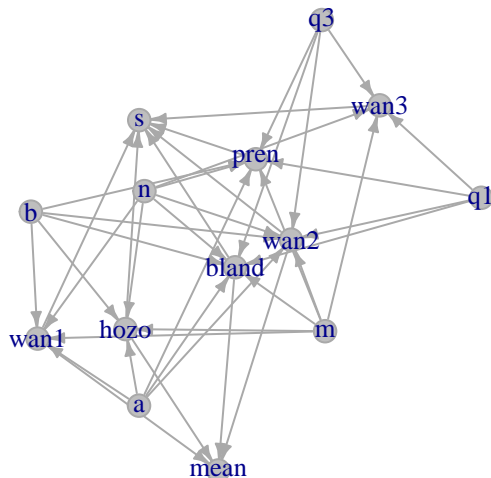
An image or diagram is very help, if not burdonsomely time consuming (but it often is).

For those comfortable with *graphs* understood as mathematical objects as a set of vertices from V , and a set of edges $V \times V$. There are visualisation optionss where the vertices can be tagged with attributes.

```
library(igraph)

confusing_relationships <- make_graph(~
a+hozo, b+hozo, m+hozo, n+hozo, hozo+s, hozo+mean,
a+wan1, b+wan1, m+wan1, n+wan1, wan1+s, wan1+mean,
a+pren, b+pren, m+pren, n+pren, pren+s,
q1+wan3, q3+wan3, m+wan3, n+wan3, wan3+s, wan3+mean,
q1+pren, q3+pren, m+pren, n+pren, pren+s,
a+wan2, b+wan2, q1+wan2, q3+wan2, m+wan2, n+wan2, wan2+s, wan2+mean,
a+bland, b+bland, q1+bland, q3+bland, m+bland, n+bland, bland+s, bland+mean
)

confusing_relationships %>% plot(
  vertex.color = "grey",
  vertex.frame.color = "darkgrey",
  vertex.size = 10,
  vertex.label.cex = 0.8,
  edge.arrow.size = 0.4
)
```



- ggplot picture of algorithm? use vis thingy from drake?

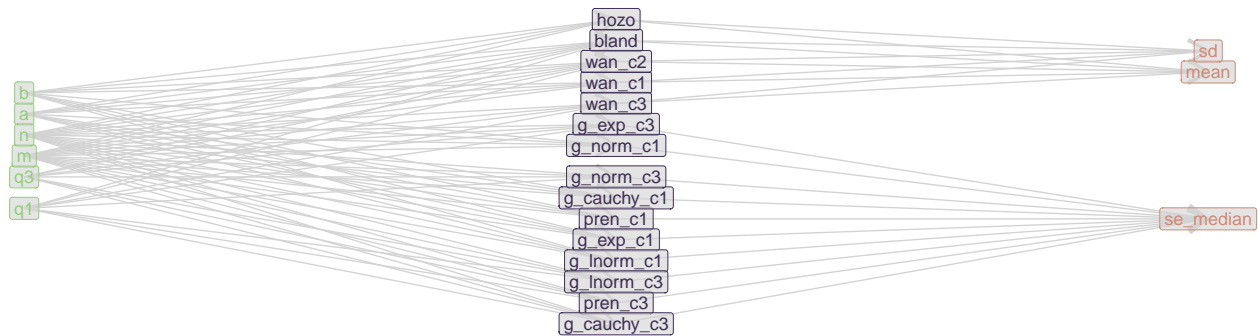


Figure 1: Codebrain before coderegistration.

Test-driven workflow

```
# image thingy
```

neet test the inputs and outputs

neet test all input cases

complete any other tests and document them

Assessing the doneness of software

Further directions

References

- Bryan, Jennifer. 2017. “Excuse Me, Do You Have a Moment to Talk About Version Control?” *PeerJ PrePrints* 5: e3159.
- Chang, Chuan C. 2009. *Fundamentals of Piano Practice*.
- Fraser, Hannah, Tim Parker, Shinichi Nakagawa, Ashley Barnett, and Fiona Fidler. 2018. “Questionable Research Practices in Ecology and Evolution.” *PLOS ONE* 13 (7): e0200303. <https://doi.org/10.1371/journal.pone.0200303>.
- Galway, James. 1990. *Flute*. Kahn & Averill.
- Gray, Charles T. 2019. “Code::Proof: Prepare for Most Weather Conditions.” In *Statistics and Data Science*, edited by Hien Nguyen, 22–41. Communications in Computer and Information Science. Singapore: Springer. https://doi.org/10.1007/978-981-15-1960-4_2.
- Hayes, Alex. 2019. “Testing Statistical Software - Aleatoric.” Blog.
- Hester, Jim. 2016. “Covr: Bringing Test Coverage to R.”
- Karp, Paul. 2019. “Robodebt: The Federal Court Ruling and What It Means for Targeted Welfare Recipients.” *The Guardian*, November.

- Marwick, Ben, Carl Boettiger, and Lincoln Mullen. 2018. “Packaging Data Analytical Work Reproducibly Using R (and Friends).” e3192v2. PeerJ Inc. <https://doi.org/10.7287/peerj.preprints.3192v2>.
- McIlroy, T. 2017. “20,000 People Sent Centrelink ‘Robo-Debt’ Notices Found to Owe Less or Nothing.” *Canberra Times* 13.
- Riederer, Emily. 2019. “RMarkdown Driven Development (RmdDD).” *Emily Riederer*.
- “RStudio Cloud.” n.d. <https://rstudio.cloud/learn/primers>.
- Wickham, H. 2015. *R Packages: Organize, Test, Document, and Share Your Code*. O’Reilly Media.
- Wilson, Greg, D. A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven H. D. Haddock, et al. 2014. “Best Practices for Scientific Computing.” Edited by Jonathan A. Eisen. *PLoS Biology* 12 (1): e1001745. <https://doi.org/10.1371/journal.pbio.1001745>.
- Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. 2017. “Good Enough Practices in Scientific Computing.” Edited by Francis Ouellette. *PLOS Computational Biology* 13 (6): e1005510. <https://doi.org/10.1371/journal.pcbi.1005510>.
- Yuan, Ding, Yu Luo, Xin Zhuang, Guilherme Renna Rodrigues, Xu Zhao, Yongle Zhang, Pranay U. Jain, and Michael Stumm. 2014. “Simple Testing Can Prevent Most Critical Failures: An Analysis of Production Failures in Distributed Data-Intensive Systems.” In *Proceedings of the 11th USENIX Conference on Operating Systems Design and Implementation*, 249–65. OSDI’14. Broomfield, CO: USENIX Association.