COMPJ-2022-05-0325: Response to referees

Harold Thimbleby

May 11, 2023

Thank you for this opportunity to respond to reviewers' comments. The reviewers' comments and queries, which I have really appreciated, have resulted in many helpful improvements to the paper, as detailed below.

I've also reread the entire paper and taken this opportunity to improve the language and clarity throughout.

I have written my responses to the reviewers' comments in red below. If it helps in referring to them in future correspondence, I have numbered reviewers' points and my specific responses.

1 REVIEWERS' COMMENTS:

1.1 Reviewer: 1

(**1**) I find this paper to be much improved from the initial submission, and it (and the author's comments) have convinced me that it is valuable to publish, at least to spark more discussion of the issues it raises.

1 | Thanks for this positive feedback!

However, I still think there are some concerns that should be resolved, as detailed below.

(**2**) page 1, lines 12-16 — slight changes in text Scientists rarely ensure the structure and quality of the code they rely on, and they rarely make it accessible for scrutiny. Even if available, scientists rarely provide adequate documentation to understand or use their code reliably. When code used in science avoids adequate scrutiny, published science will be unreliable, because results — data, graphs, images, inferences, etc — generated by the code will themselves be unreliable.

2 | Suggested corrections made.

(**3**) page 1, line 17 — slight changes in text This paper proposes and justifies...

3 | Yes! I like the improved order

(**4**) The idea of a table of contents in the body of the paper seems odd, but I assume this will be taken care of in the journal production process

4 | I left the table of contents. If it is not required in the Journal style, it's
  | easier to delete than to reinstate.

(**5**) page 2, column 1, line 40 I don't believe that code is generally incorrect, at least not to point of leading to incorrect results. I accept the code is generally informal and inaccessible, and that code might have bugs, but many bugs don't lead to significantly incorrect results. (The study in Table 2, while valuable, doesn't discuss incorrect code, and is also quite small. (Note that a large-scale study of incorrect code would be valuable.) The studies mentioned in 2.a [15-16] do not study the quality of code itself, but rather reproducibility of results more broadly, including issues related to finding correct versions of dependencies, input files, platforms, etc.; this is a longer discussion than belongs in this review.

5 | I've re-instated a brief "further work" section that mentions these issues
  | for further work. I've also added citations to some papers that discuss
  | automatically finding dependencies etc.

(**6**) page 4, column 2, lines 11-13 — slight change in text: While chemists are trained in reliable methodologies, taking quality glassware for granted is reasonable.

6 | Thanks; fixed

(**7**) page 4, column 2, lines 42-44 references [9, 17-20, 20] doesn't make sense with 20 listed twice.

7 | Oops! Thanks; fixed

(**8**) page 4, column 2, lines 41-49 I note reviewer 2 mentioned JOSS, which I think should be mentioned here in addition to ReScience, as it specifically does focus on peer review of code, including code quality, at least to some extent, though not to the extent the author calls for.

8 | I now mention both journals.

(**9**) page 5, column 1, lines 11-13 I don't think TOP is stricter than FAIR, but rather, has a somewhat different focus. In terms of code quality, it is stricter, but in other aspects (e.g. findability), FAIR is stricter.

9 | Thannks. I've improved the wording now.

(**10**) The discussion in 2.b about computable papers is reasonable but limited to the cases where the computing is can be done on purely local resources, or freely available resources. Where more computing is needed (e.g., HPC codes, this idea becomes much more complex very quickly.) Also, I suggest that Binder be mentioned in the Jupyter bullet.

10 | Thanks, I've now mentioned HPC, Binder, etc, in a new brief section on learning curves.

(**11**) In 2.c, it's unclear how specific RAP is meant to be. Is this a general description of a process, which can be implemented by different people using different specific tools, or is this a description of a specific set of tools used in a specified manner?

11 | Table 4 (which I've improved a bit to help address this concern), in particular, makes it clear that RAP is a generic workflow.

(**12**) Table 5 isn't really a table — again, the journal should specific what this kind of object should be called.

12 | I've now made the contents look more like a table! I've also ensured all tables have a consistent design.

(**13**) page 10, column 2, line 48 Something is wrong with the grammar in "used comment to inactivate code,"

13 | Thanks; fixed

(**14**) Regarding the note about section 3.a on page 11, what happens if the company that sells Mathematic goes out of business? How would this notebook then be usable in the future? This also applied to the example about [33] in 2.b. What if the author had chosen a technology in 1999 that was dropped by its commercial vendor in 2001? How likely is it that there would be a way to run that notebook 22 years later?

14 | These are important points! I have addressed them in a footnote in the earlier section, 2.b.

(**15**) page 13, column 1, line 24 Please explain what the Code of Hammurabi has to do with building houses

15 | Short footnote added to explain

(16) page 13, column 2, lines 28 data and code are not equivalent or interchangeable. For example, one is creative whole the other is not. This is like saying numbers and literature are equivalent. While I completely agree that there should be policies on code, and that these policies should be strict, it's not because code and data are the same, but because both are important to research.

16 | Thanks for this observation. I've rewritten the section, explaining the issues more clearly, and making the argument more evidence-based and hence less provocative.

(17) I feel that some of sections 4.b and 4.c don't really match their titles, nor how they examples are divided. For example, the last paragraph of 4.b is about the status of code in publications, not about the status of coding.

17 | I've deleted the last paragraph of 4.b.

(18) And much of 4.c is about PRISMA and how papers change over versions, not really about coding.

18 | I've rewritten 4.c to make it clear that PRISMA is a leading example: the section is not about PRISMA, but about publication policies and standards, of which PRISMA is a leading (and up to date) example.

(19) Supplemental material, page 41, line 49 — have data policies that are weaker should be have code policies that are weaker.

19 | Thanks; fixed

(20) In 4.d, saying that code and data are formally equivalent (and see comment above) only makes sense in a limited context (that discussed in the supplementary material), but this is overstated here.

20 | See point 16 and my response to it above, which I think is the same point.

(21) In 4.d, "bugs naive bugs" has an extra "bugs"

21 | Thanks; fixed

(22) In 4.d, the last paragraph's discussion of intentional bugs is confusing terminology. It would be much cleaner to talk about bugs in translating from theory to model and in translating from models to code. I don't think that either kind is actually intentional.

22 | I've rewritten this paragraph to make it clearer; it also cites material that provides further discussion of the term used.

(23) page 17, column 1, line 37 this is only surprising in the sense that gambling in Rick's Cafe was shocking

23 | I think it's surprising that I only say surprising once in the entire paper . . .

(24) I am still unhappy with the use of "repository" in 5.d to refer to both Dryad and GitHub, since one is an archival system that requires metadata and involves curation, while the other has none of these properties, and is a social coding platform build on top of a version control system.

4

24 | Dryad and Github certainly differ in detail, but both describe themselves as repositories — indeed Dryad started out life as DRIADE, The Digital **Repository** of Information and Data for Evolution.

(**25**) The next paragraph talks about using the repository for maintaining the code, which cannot be done in Dryad, but could be done in GitHub.

25 | I've added a footnote: "Dryad `datadryad.org` curates raw, unprocessed data. At the time of writing, Dryad excludes code; however, it uses a separate organization, Zenodo `zenodo.org`, to host code and other relevant information. This arbitrary separation is unfortunate as it increases management problems, increases reproducibility problems, limits using RAP, and most seriously limits how scientists can structure their data and code to best suit their research (see section 4.d)."

(**26**) Note that the final paragraph of this section talks about differences between code and data, which are correct but seems to make the earlier comments about these being formally equivalent incorrect without more discussion earlier.

26 | Thanks to your ideas, I improved the earlier discussion, and think this fully addresses this point.

(**27**) The abrupt change of topic that comes with section 6 is confusing. I would prefer to have seen some general discussion of what the author thinks is needed, with RAP+ then discussed as a way to implement this, rather than it being pitched as The Solution. The discussion in Section 7 seems to disconnect code from the resources used to run the code, and assumes that running code is simple and has no cost. This carries over into section 8 as well. While section 8 is the call to action, this is really just a continuation of the more opinion/vision part of the paper that started in section 6. Some restructuring or changed titles might be useful here. Section 6 could be future vision, with the contents of 6, 7, and 8 being subsections.

27 | I've rearranged and grouped all these sections under a new main heading "Rethinking science that uses code." (The table of contents on page 2 makes clear what I've done.) I've also done minor rewording to improve the flow to be consistent with the new improved structure. Thanks for this observation.

1.2 Reviewer: 2

The revised manuscript has improved in many ways. In addition to addressing many detailed comments by both reviewers, it is now clearly an opinion article. I still find it a bit lengthy and lacking actionable proposals that could initiate a transition towards the better future described by the author.

Two minor comments:

(**28**) Page 20, line 16: Almost the entire scientific process can be automated... If "scientific process" means all of science, then this is clearly not true. Specific protocols for data acquisition, data analysis, etc. can be automated, but not all of science! Otherwise... I

would happily let computers review manuscripts instead of doing it myself!

28 | good point! thanks; fixed. I've also replaced the word "process" (where relevant) with "workflow" as this word is more specific, without including "thinking" that may be implied by "process."

(29) Page 20, line 40: To date, this critical point has been overlooked.

29 | (Not sure what the point is here — this is what the paper originally said.)

(30) While this is true for the majority of computational scientists, there is some prior work in this space. See for example Maneage (https://maneage.org/), or the work by Courtès (https://rescience.github.io/bibliography/Courtes_2020.html) that extends the Guix package manager. Both approaches are very similar to the RAP* proposal.

30 | Thanks for these exciting references. I added a citation to the impressive system Maneage, and mentioned that it also includes a good review in its appendix. I've also added a couple of paragraphs about Courtès's impressive work in the section on RAP.

A few typos I noticed:

(31) - Page 14, line 12: bugs naïive bugs ⇒ naïve bugs

31 | Thanks; fixed

(32) - Page 16, line 46: Ferguson's reported science unlikely to be reliable. (insert "is")

32 | Thanks; fixed

(33) - Page 41, line 49:. . . some journals have data policies that are weaker than their data policies. . . ⇒ code policies

33 | Thanks; fixed

(34) Konrad Hinsen, CNRS, France (Note to editor: I have decided to sign all my reviews. Please don't remove my signature.)

34 | Thanks so much Konrad. I've really appreciated our correspondence, your critical checking, and you pushing me to improve the reproducibility of my own work. Thanks.

1.3   Reviewer: 3

(35) There are a few remaining typos

35 | Thanks. I've reviewed the entire paper very carefully, and of course benefitted from all the referees' careful reading and pointing out of typos, all of which I have fixed.

**** . . . including one sentence where you may have omitted a "not", thereby reversing what you intended to say. (I regret that I have been unable to find this again — please check. I think it was toward the bottom of a paragraph in the left-hand column.)

(36) On p16 column 1, the sentence starting "Alternatively) near line 13 lacks a verb.

36 | Thanks; fixed

(37) The sentence starting "Ferguson" at line 46 omits "is".

37 | Thanks; fixed