Based on my own evaluation and that of the two reviewers, I believe this paper is important and would contribute greatly to the discussion on experimental reproducibility and replicability in science. The reviewers have raised several points that should be addressed. Most of the comments suggest re-arranging the structure of the paper. Reviewer one highlight the over-representation of examples associated with microbiome characterization, and has suggested to either equilibrate the balance and include other example or refocus the article on microbiome research. I think that either would be fine. Please address the comments in a point-by-point response explaining how you have modified the manuscript. Also, please provide a "track-change" version of your revised manuscript so it can be easily re-assessed.

- Have put focus on microbiome research and emphasizing the difficulty of making our work reproducible

Reviewer #1 (Comments for the Author): This paper is a bit of a mix of a framework for reproducibility and replicability, criticism of human microbiome studies through examples, a response to an AAM report, some of the author's own approaches in their lab and exercises where people can evaluate their perspective on some of these issues. The definitions and framework in Table 1 are great, as are the Exercises and the general message around the challenge of reproducibility and replication. However, there's not a clear narrative throughout the paper. Upon reading the title, my first impression that this would be a paper about how it is difficult to do the work that makes research replicable and reproducible, and looking at the different aspects of that. I think that message comes through, but is mixed in with examples of studies that might not be relevant to those points, and the topics jump around a bit. This message around the challenges of reproducibility is an important one, and the overall idea of this paper is good. Some rearranging and clarity in narrative would be needed for this paper to be more effective. I admit I like the more general message, but given that there are many examples from the human microbiome, one option would be for the paper to focus on reproducibility in human microbiome work.

- Have put focus on microbiome research and emphasizing the difficulty of making our work reproducible

While the title is catchy, I think it could be something else that captures a similar nuance without the !&* (I know, I'm stodgy)

In the beginning there is a mention of the AAM report, and it seems like this paper is partly a response to that report. However, that point comes in a little later, and it's not clear what points the AAM report is making that are being responded to. Not sure if more or less of that should be

included, and that would depend on the general narrative.

It's great that reproducibility and replicability are defined, but they are used extensively before they are defined. It would be good to move those definitions sooner.

I like Table 1! Clear way to lay out the differences and like attention to robustness and generalizability as well.

Thank you! I am glad to see it was helpful for creating a framework for thinking about how we do research.

The 'example' section is a good example to use to discuss the different concepts, but in several sections it highlights poor methodology, underpowered experimental design, and poorly described data curation methods. Both the tone and the idea of a poorly designed or executed experiment distract from the point about reproducibility. That paragraph also lacks clarity overall. Is the author's point that this was just a bad study from a "this isn't how we do science" perspective, or that there was good intent, but reproducibility/replicablity/robustness are difficult (which one might expect from the title of the paper) and this study didn't meet those challenges. Whether or not true, this paragraph comes off more as a attack on that study than a broader statement on reproducibility. I think this section can be rewritten with this same example, or there might be better examples. In signaling out a single study to make a broader point, the attention often ends up on that single study rather than the bigger issue.

In the 'Reproducibility' section there are good points and sets of recommendations, but topics jump around. It goes from workflow to raw data, to technology and software updates and on. This section could use a clearer narrative or subsections on these different components.

In the 'Replicability' section, I'm not sure that the first example is a problem with replicability or as the author suggests, that this is an issue with experimental design. An issue with replicability, as defined by the author, would not be that they were trying to compare geographies, but that they were studying some question or effect and seeing if it took place in both geographies. From that perspective, it may have been a good example of replicability. It does highlight the challenges that sequencing at different centers or having different people do sampling can be confounding, an important issue in replicability. So, perhaps that could be highlighted. Still a better example might be one where there were two planned separated studies of the same thing and the result did not replicate. The second paragraph describes some examples well.

I'm not sure the section on statistical significance is a replicability issue either. It's not that things aren't replicating because different systems are being analyzed, but because statistics are being applied incorrectly. Maybe the overall take home of this section is that often things don't replicate, because something wasn't done correctly, but there are lots of reasons, and ones often hard to detect, why something wouldn't be done 'correctly'. This is an important and valid point, but is different from the one that replicability is difficult because of real biological variation in populations and individuals. Perhaps it would be possible to break those out more in this section.

In the 'Robustness' section, the point about applying multiple techniques is good. In the intro, it was also mentioned about performing multiple types of computational analyses, for instance using different software intended to assess the same things. This would be a good thing to include in this section as it is something more easily and more frequently done, and helps with robustness of the quantitative analysis.

The exercises are great, but they don't lead naturally from the previous sections. I strongly agree with the idea that a culture of reproducibility is important, but it hadn't been mentioned previously in the paper to well motivate why these exercises might be helpful. One idea is that exercises could be tied to the different concepts of reproducibility, replicability, robustness and generalizability.

The author's Riffomonas project is very relevant to this topic. Perhaps there could be some more on that project in the paper. One other narrative approach could be that reproducibility is hard, for these general sets of reasons, and these are some resources to help, particularly in microbiology, with the Riffomonas project being one of those resources, and having the opportunity to add some more information on that project.

Reviewer #2 (Comments for the Author):

General comments * Overall I like the article and think it will be of broad use * It probably would be helpful to better describe a workflow involving sequencing and also better describe methods used when they come up so that this paper has broader possible impact. For example you could describe metagenomic sequencing, rRNA gene sequencing, etc.

Abstract: Re "We need to respect that science and microbiology, in particular, are difficult." I am not convinced that there is justification to say microbiology is difficult

I do not think there is support for the "Of course" part of this statement: "Of course, Leeuwenhoek's

work was rigorous, impactful, and robust." Why couldn't this be non rigorous? What do you assume otherwise?

This sentence has been edited.

Same type of comment with "Few would suggest that Bourne's group was sloppy or that they failed to be transparent." Why? I mean, I think Phil Bourne is a great scientist but what evidence is there that few would suggest this.

"This Perspective will use the most widely used definitions" - do you have any information that shows that these are the most widely used?

"Although each part of their approach had significant weaknesses including methodological biases and underpowered experimental designs". Is this shown in the paper about this? Or is this a new comment? If new, please explain further. If from the paper about this please add citation to this sentence.

Re "However, researchers still fail to post their sequencing data to public databases or do not provide the necessary metadata with the sequencing data" I assume you mean "some researchers" for this part and should say that.

"Because many journals impose word limits on manuscripts, Materials and Methods sections become a chain of citations to previous work that each cite previous work (10)." I agree with the second part of this and also agree that limits on words are annoying. But is there evidence that this issue is DUE to the limits on words?

Re "For example, we used GNU Make to write a workflow in our meta-analysis of the obesity data, such that downloading a copy of the scripts from the project's GitHub repository and writing"make write.paper" in the command line will reproduce our analysis" This is a great thing. However, the way it is written it sort of implies that simply by using GNU Make this will happen. Instead it is important to TEST whether this worked and not just assume that using GNU Make will allow others to perfectly reproduce something.

Re "Unfortunately, the reliance on web-based workflows like GenBank (https://www.ncbi.nlm.nih.gov/genbank), greengenes, RDP, and SILVA preclude analyzing new data with older versions of the sites." The wording here seems off. These are not workflows but web sites that have various workflows available at them. Can you revise?

Re "are likely to persist for at least a decade." Why a decade?

Re "Combined with version control software such as git, these literate programming documents can allow researchers to document and share the evolution of their analyses." Can you elaborate on how someone might use such notebooks to explore p hacking or the Garden of Many Forking paths. Has anyone actually done this with such notebooks? It sounds nice but also could be very hard for someone to use such notebooks in this way.

Minor thing: "For example, the Human Microbiome Project used Roche's 454 platform to sequence the 16S rRNA gene (23)." should probably say "16S rRNA genes"

Typo in "Washington University sequenced the DNA from the St. Louis subject" should be subjects I assume.

I doubt this statement is true "Forgotten in discussions of replication failures". I have heard this discussed in many conferences / talks focused on scientific practice

This statement is not quite true: "Metabolomic data would be even more independent from the DNA-based methods since it requires completely different sample handling and processing steps" They frequently overlap in some steps.

Re "The gold standard of science is to have a result that is generalizable across populations". I believe this is just not true. First, the term "populations" only applies to some types of research and thus this only could in theory apply to some fields. Second, even when fields have populations I do not think this is always a gold standard. One can conduct interesting science on single samples without any need for application across populations. I would recommend removing or rewriting this statement.

Re "Yet, many microbiology training programs focus on laboratory skills while ignoring data analysis skills." and data on this would be very useful. Certainly we need more data analysis training but I am not sure there are many micro training programs these days that completely ignore data analysis.

In the beginning section on the paper airplane it would be good to have some more explanation of what the questions are for. That is, after "Have the participants trade instructions, separate, and implement the instructions." you have a bunch of questions. Are these for students to answer individually? In groups? Are these for the instructor to answer? How would you make use of these

questions?

In #2 a similar issue. Who are the questions for? Who is supposed to answer them? Is this for a graduate student or are more senior person who wrote the recent paper and just that they are supposed to imagine how a grad student would work? I am confused as to how one would use this scenario.

In #3 I think I get who the questions are for.

For the 1st paragraph, I confess to being a bit confused here and I think this is partly that this is not really a conclusion but a new exercise of sorts. I do not grok what you are getting at with one of the collaborators being "themselves six months ago, and old them does not have email access". What does email have to do with this? Why six months? How is this a collaborator.

Then I also do not really get "Their second most important collaborator is the director of their research group." Why? Why does the director of a group have such prominence here? I have many students in my lab who work on independent projects that have nothing to do with me and I am not an author on these papers. Yet I am the director of their research group. The prominence of the director of a research group in this conclusion confuses me.

I know I can be dense but I am going to guess that other people might be confused here. Please expand this section and make the transition to this conceptual exercise smoother and better laid out.

Some reviewers may not like "Although there is much room for improvement, we must acknowledge that science is a process of learning and that it is really f#$%ing hard.". But I do

In Table 2 I don't like the use of the 1st person "I". In many cases it should be "we". Not sure there is a solution to this but just thought I would point out that "I" is a bit inaccurate.

I think it would be good to add to the legend "These are example questions on might ask for this category." You could even say "Add your own questions too"

Regarding sex I am not sure if that is the ideal term for "participants' since some studies focus on gender not sex.

While I appreciate phylogeny more than most I don't think this term is ideal: monophyletic folder structure

Regarding tools you have "Are free tools used in preference to proprietary commercial tools?." What about openness and not just "freeness"?

Re reagents "Do I include a table of the cell lines, strains, genotypes, and primer sequences that I used?" might be better to have ""Do I include a table of reagents such as cell lines, strains, genotypes, and primer sequences that I used¿'