Reproducibility, Replicability, and Reliability

Xiao-Li Meng^{1,2}

¹Department of Statistics, Faculty of Arts and Sciences, Harvard University, Cambridge, Massachusetts, United States of America,

²Harvard Data Science Review, Harvard Data Science Initiative, Harvard University, Cambridge, Massachusetts, United States of America

Published on: Oct 29, 2020

DOI: <https://doi.org/10.1162/99608f92.dbfce7f9>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

When I was invited to be a moderator for a panel on the U.S. National Academies of Sciences, Engineering and Medicine (NASEM) 2019 report, “[Reproducibility and Replicability in Science](#)” (hereafter “R&R”), I got a bit nervous. No, I was not worried about my research failing on one R or both. (This, of course, is not a bragging line, but a bottom line for any researcher with self-respect.) I was nervous because I knew the terms *reproducibility* and *replicability* had been used interchangeably by some communities, and with very different meanings by others. I had my own working definitions, of course. But would they be consistent with the ones in R&R?

It was therefore a big relief when I saw R&R equated reproducibility with *computational reproducibility*, and associated replicability with *scientific replicability*, as clearly spelled out in the [conversation](#) with Harvey Fineberg and Victoria Stodden in this issue of *HDSR*. As a statistical researcher and educator, and as the Editor-in-Chief of *HDSR*, I am deeply grateful to Harvey for chairing the NASEM’s committee that produced this timely and critical report, and to Victoria not only for her role on the committee but most importantly for being the editor of this special theme on reproducibility and replicability. Victoria’s editorial (forthcoming) provides a succinct and insightful introduction to each article in the special theme, which permits my editorial to focus on the general interplay among reproducibility, replicability, and data science.

Below I would like to invite interested readers to contemplate with me a few momentous challenges we face as a data science community, both as an intellectual exercise and as mental preparation for diving into this rich collection of reactions to, reflections on and research for R&R. Of course, the 234-page R&R report contains far more material than I possibly can touch upon, especially with my very limited perspectives and experiences as a lone statistician. I also doubt many readers would have the time to go through the entire R&R. But I’d highly recommend reading, minimally, its [Summary](#) (14 pages). I also find Section 2, “Scientific Methods and Knowledge” (11 pages), an extremely rewarding read, regardless of whether you are a senior scientist or a senior in high school.

Similarly, I am deeply grateful to the editors of another special theme in this issue, predicting the 2020 U.S. Elections. Liberty Vittert, Ryan Enos, and Stephen Ansolabehere provided their [lively introduction](#) to each of the articles in the election theme (and bravely offered their own predictions), which allows me to elaborate on the ultimate reason for addressing reproducibility and replicability. That is, to ensure the *reliability* of our data-driven findings, for scientific understanding, daily life guidelines, political forecasting, or any of the other reasons for which we conduct research. Our collective failure in reliably predicting the 2016 US election should induce a painful memory for anyone who cares about data science done right, regardless of political inclination. Would we do (much) better this time? The answer is just around the corner, but I hope our pre-released (October 27, 2020) election compilation has helped or will help you better anticipate the election outcome than in 2016.

As we are living in a devastatingly volatile time, our ability to acquire and maintain reliable information can become a matter of life and death, literally and figuratively. Even the basic statistical literacy can have that effect, as Carlo Rovelli wrote, [“Statistical illiteracy isn't a niche problem. During a pandemic, it can be fatal.”](#) The triple threat of producing, consuming and disseminating biased information and misinformation, brought by 2020 to an all-time high, makes it ever more urgent to double our efforts at the education front. Effective, early, and unprejudiced education is an ultimate and sustainable force in curtailing any tendency of prioritizing self-serving over fact-serving, and in enhancing our ability to engage each other with warm hearts and cool heads. Early general education on quantitative literacy and reasoning under uncertainty is the key to ensure reliable production and consumption of scientific findings in both short and long runs. As part of this on-going effort, *HDSR* features in this issue (yet) another three educational articles, from high school to two-year colleges, highlighted towards the end of this editorial.

Of course, even—or especially—at the most difficult times, life offers more to the diligent seekers of its joy. I therefore cannot possibly end this editorial without reporting whether my happy [pairing of okra and *HDSR*](#) has been reproduced or replicated by our readers. So you have been warned (or, rather, your appetite has been whetted).

Reproducibility

The interplay between data science and reproducibility should become immediate transparent when R&R synonymizes the term reproducibility with *computational reproducibility*. Reproducibility comes before replicability because it is a minimal requirement for any reported results to be trusted. Findings may or may not actually be valid, but it would be very hard to convince anyone if we cannot even be self-consistent, that is, delivering our products as we advertise. Whatever results we compute and report, someone else who uses exactly the same data, methods or algorithms should be able to reproduce the same results. Otherwise we have some explaining to do, minimally to ourselves. That is, in a layperson's term, the concept of reproducibility is as simple as saying that this basket contains a dozen eggs, then whoever counts it, there should be exactly 12 eggs (unless the counter is arithmetically challenged).

As a lesser toy example, suppose we estimate from a data set that a 95% confidence interval for a disease prevalence in a population is 8 to 12%. We may never know if that specific interval covers the actual rate, but the interval (0.08, 0.12) should still be the output if someone else implements exactly our procedure with our data, at least when the procedure is a deterministic one. Furthermore, suppose we justify our confidence procedure via a simulation study, which shows that the procedure covers the truth about 95% of the time under some idealized simulation configuration. Then someone else who carries out exactly the same simulation study should also obtain the same 95% coverage, barring acceptable Monte Carlo errors (which can also be eliminated if we have reported the random seed for our simulation study).

However, readers who have had some experience with checking reproducibility are likely to tell the rest of us that it is far easier said than done. First, in order for others to verify the results we report, we need to report all the ingredients needed and make them accessible. Right there, we have a big problem. We need to provide data, but we might not be willing, which is not always for a self-serving purpose; e.g., we may want to protect data proprietary for our under-resourced collaborators out of fairness consideration for their investment. We may not even be allowed to disclose our data because their privacy is protected by law or by confidentiality agreements. We may need to provide code, but again we may not be willing or are not allowed to do so for similar reasons (e.g., proprietary constraints). That is, reproducibility is strongly tied to transparency and open science/access, as emphasized in R&R, both being complex topics that involve multiple disciplines and stakeholders, from computer science to library science, and from publication industries to funding agencies.

Secondly, I have kept saying ‘someone else’ is charged to check on what we report. But *who* are they? Who has the incentive or time to check on all our data, algorithms, implementations, etc.? For any of us who have done it, we know how time-consuming it is to verify other people’s work, even if other people fully collaborate. Although occasionally debunking others’ mistakes can bring fame and funds, the vast majority of time it only brings frustration and foes. As much as we believe and hope that scientific integrity is a sufficient motive, the reality is that the vast majority of us who are qualified to carry out the reproducibility checks are already struggling to have adequate amount of time to conduct and triple check our own research findings to ensure their reproducibility and replicability.

Currently, the entire scientific enterprise does not have a viable professional community of ‘reproducibility reinforcers’ who are trained and rewarded for verifying other people’s research findings. The notion of science being kept clean by ‘self-policing’ is apparently not working as well as we had hoped, prompting collective efforts such as those that had led to various R&R-related initiatives (and this special theme).

Of course, some may argue that R&R highlights exactly how ‘self-correction’ works, for it is a product of the scientific community itself. This is indeed very true, and R&R is likely to have lasting impact on scientific policies, funding priorities, professional guidelines, etc. However, Marcia McNutt, President of National Academy of Sciences, argued in her (forthcoming) reflection on R&R (“[Self-Correction by Design](#)”), “we need to adopt an enterprise-wide approach to self-correction that is built into the design of science.”

I wholeheartedly support her call, and for the need to enforce reproducibility on a grand scale I’d even venture that it will need a scientific profession devoted to it, no less than we need a full workforce of cryptographic experts to ensure our cybersecurity. I will therefore leave it to readers to judge, based on their own data and information provided by this thematic issue, whether it is time to take some bold steps forward, such as establishing a ‘Department of Computational Reproducibility’ inside a school of Data Science, especially as increasingly more of the latter are being built. The article by Craig Willis and Victoria Stodden¹ (forthcoming), “[Trust but Verify: How to Leverage Policies, Workflows, and Infrastructure to Ensure Computational](#)

[Reproducibility in Publication](#),” for example, would provide excellent dissertation materials for such a department.

I am, however, not suggesting any kind of ‘Department of Replicability’ for essentially the same reason I [argue](#) that it is unwise to establish a “Department of Data Science.” Whereas reproducibility is largely a computational issue, replicability essentially is everyone’s issue, as made clear in R&R and briefly reviewed below, and hence it requires expertise and skills from all contributing fields of (data) science.

Replicability

If ensuring reproducibility is hard, then ensuring replicability is NP-hard. It is NP-hard because the metric for declaring replicability is *not possible* to be made meaningful without diving into the pertaining studies. Very strategically, R&R adopted the same “obtaining consistent results” as a general metric in their definitions of both reproducibility and replicability. This was a strategic choice because the phrase ‘consistent,’ ironically, does not have a consistent meaning. For computational reproducibility, consistency typically takes on a rather stringent meaning. If I ask two students to independently verify a p-value calculation using the same test procedure and the same data, and they report respectively that $p=0.012$ and $p=0.021$, I’d know immediately that at least one of these numbers is wrong, and we will need to find out what went wrong. Coding errors? Numerical errors due to different approximations? Reporting errors, as the two digits accidentally got switched? You would not find me excited: ‘Oh, great, the results are confirmed because both are significant at 0.05—another publication!’

On the other hand, in the context of replicability, we may indeed congratulate ourselves for having obtained similarly significant results when these two p-values are from two independent studies designed to test the same hypothesis. Anyone proposes identical p-values as evidence for replicability would immediately reveal themselves as being statistically challenged. Indeed, (nearly) identical numerical results in the context of replicability may even raise the suspicion that someone had cooked the books.

Does this mean that it is actually easier to verify replicability, since we can give ourselves some leeway in permitting differences? Not at all, since it is this very ‘leeway’ that opens a troubling path to the evasive world of *scientific replicability*. As R&R stated, “Two studies may be considered to have replicated if they obtain consistent results given the level of uncertainty inherent in the system under study.” A spot-on definition, since surely we must allow some variations arise when we move from one study to another, for example, because of sampling variability. And surely we should also allow for some variations for the team differences. Even if two teams work on the same data, insisting that they produce the same results may do some serious harm to scientific progress. I was reminded of such phenomena years ago by a friend who was serving as a statistical consultant to a financial firm. The firm had three consultants to work on the same project, and it insisted that the three must replicate each others’ results before their results could be accepted. Before I had a chance to praise the firm’s statistical integrity, my friend lamented that this seemingly desirable replicability had led to

the most stupid analysis one can think of for the project because that was the only method that all three consultants could understand, implement, and verify.

However, to set up a tolerance level for acceptable non-probabilistic variations is far more complicated than for probabilistic ones, even if the former are quantifiable. These include (at least) variations in data quality, modelling strategies or experiences, and prior or substantive knowledge. And which parts of these variations should be counted towards signal and which parts towards noise?

To make the matter even more complicated, while reproducibility is always desirable, replicability is not. As R&R pointed out, scientific progresses are made by having studies that improve upon previous ones, or do things that could not be done before. Harvey makes clear in his [interview](#) that a part of the controversy about the notion of replicability is that “[s]ome see the failure to replicate as a source of potential insight and deeper understanding, while others see it as indicating a lack of rigor in science.” There are, therefore, “helpful and unhelpful sources of non-replication.” Judging what is helpful or not helpful clearly will be problem-dependent and requires judicious judgment based on substantive knowledge.

Designing sound replication studies therefore requires a host of data science skills, from statistical designs to causal inference to signal-noise separation, that are simultaneously tailored by and aimed at substantive understanding. Causal inference is critical here since we surely do not want to blindly replicate confounding factors to trap ourselves by the replication paradox: *the more we replicate, the surer we fool ourselves*. Our collective over-confidence in the predictions for 2016 U.S. presidential election precisely was built upon seeing biased poll results replicated (too) many times. This example should also remind us that replicability cannot possibly be an aim by itself for ensuring scientific integrity.

Reliability

The overall aim of reproducibility and replicability, of course, is to ensure that our (research) findings are reliable. Reliability does not imply absolute truth, which is an epistemologically debatable notion to start with. But reliability *does* require that our findings can be triangulated, that is, they are reasonably robust to the (relevant) data or methods we employ. Furthermore, they need to pass reasonable stress tests, such as out-of-sample predictions, fair-minded sensitivity tests, and a lack of serious contradictions with the best available theory or scientific consensus, unless the findings are designed to challenge the existing common wisdom.

Reliably predicting election outcome is perhaps the data scientific endeavor most demanded by the general public. Opinion polling is the most common method employed by social scientists, the media, and others to give us a peek into the future (thought that’s not the only tool, as we are reminded by Allan Lichtman’s “[Keys to the White House](#),” a qualitative predictive method that has had reasonably reliable predictions since 1984). Although few people would treat opinion polls as crystal balls, we do all more or less have some belief in their reliability when we pay attention to them. We typically measure the reliability of a pollster by its track record: how often did it get it right? A great example is given in the [article by Williams et. al.](#), which describes

Øptimus’s predictive models for presidential and congressional general elections in collaboration with Decision Desk HQ. Although readers are not given all the ingredients of their models due to proprietorial information, readers can gain reasonable confidence (with all the due caveats that models are only as good as the data they fit) in their models’ reliability from the reported performance of these models in the 2018 U.S. congressional elections, where their accuracies were largely over 90%.

Such a common-sense metric is also very useful when we build a new method via so-called backtesting: How well it would have done if it were applied to a past prediction problem? [Merlin Heidemanns, Andrew Gelman, and G. Elliott Morris](#) backtested their dynamic Bayesian model on the 2016 polling data, and found their model still overestimated the vote shares for Clinton in key Midwestern states. I am grateful to them for providing a good illustration of the efforts that can be made by authors to gain the trust of the readers: conducting and reporting reliability check results regardless of how they turn out. Most readers do not expect that everything works out optimally, but they do expect to be informed honestly. The authors’ backtesting results certainly have helped me better appreciate the caveats in their abstract about their prediction that the Democratic party’s winning chance is 80-90%. Taking these caveats into account, a reader can be much better prepared for the possibility, however small, that the race next week (the week of November 3) turns out to be a much tighter one than all the now-casts have been telling us. Indeed, [Michael Isakov and Shiro Kuriwaki’s analysis](#)² provides a scenario how this could happen if the non-response bias in 2016 election and pollsters’ ability of correcting them have not changed much. Their analysis therefore provides a stress test on the reliability of the current polling results.

Without much spoiling the readers’ joy of discovering all the treasures, let me mention that the forthcoming theme on R&R documents many more reliability-enforcing efforts by researchers from a variety of disciplines, from economics to particle physics and to climate science.

Reinforcing Early Learning

Producing reliable scientific findings takes many talents, and consuming them effectively takes even more educated citizens. Education, then, is of paramount importance, and this was well recognized in R&R. Indeed, [President McNutt’s article](#) made it crystal clear in her call to build ‘self-correction’ into the design of science: “Principles of such an approach include educating students to perform and document reproducible research, sharing information openly, validating work in new ways that matter, creating tools to make self-correction easy and natural, and fundamentally shifting the culture of science to honor rigor.” Again, I cannot possibly agree more of this loud and clear emphasis on education, and to shift the culture as early as possible.

Indeed, scientific research starts as early as in high school, as clearly reflected in [“High School Data Science Review”](#) by Angelina Chen, a high school junior, who passionately calls for a reform of data science education in high school. Angelina’s call is precisely for better preparing both the producers and consumers of (data) science, because “[s]tarting to educate kids about statistics from a younger age [...] is vital towards filling the

gap between the supply and demand of statisticians and data scientists, as well as creating a more data-literate society, both of which are necessities for the future.” This sentiment is well reflected in another “[Minding the Future](#)” column [article by Christine Franklin and Anna Bargagliotti](#), who provided an introduction to the pre-college *Guidelines for Assessment and Instruction in Statistics Education II: A Framework for Statistics and Data Science Education report* (GAISE II). The ultimate aim of this report is to “support all students to gain an appreciation for the vital role of statistical reasoning and data science, and to acquire the essential life skill of data literacy.”

To entice students to enter as early as possible the complex world of reasoning under uncertainty, applications of their interest are of great importance. In both the current and [previous](#) columns, Angelina Chen gives a number of great examples as a peer high school student. In this issue’s “[Recreations of Randomness](#),” Brian Macdonald provides a [fascinating account](#) of how statistical methods and machine learning algorithms have been used to track and analyze players’ movements on basketball courts or football fields. Sports analytics are known to trigger early interests in statistics even during the days when ‘sports statistics’ simply meant to keep track of numbers such as batting averages and winning percentages, that is, those ‘boring statistics.’ Just imagine how attractive sports analytics must have become now, with all the fancy algorithms and data visualizations. Indeed, even I got excited by seeing how the bar charts and histograms got animated in this article, even though I still don’t quite get why in the U.S., players are allowed to use hands (all the time) to play football!

As we think about early education, we should not forget early education at *college* levels, especially at two-year colleges, which typically do not get as much attention as they deserve, for they serve many more students than we often realize. I am therefore extremely grateful to Professor Brian Kotz of [Montgomery College](#), which serves 54,000 students each year, for both his broad vision for two-year college data science education and his detailed account of how they are being implemented at Montgomery College through its Data Science Certificate program, launched in the Fall of 2017. [As Brian put it](#) succinctly, “Two-year colleges are poised to play a substantial and possibly transformative role in data science and undergraduate data science education.” I therefore urge readers who are interested in data science education to take a close look at this article to learn about their successes and lessons, regardless whether you are familiar with two-year colleges or not.

Clearly all of these on-going educational effort and reforms will serve well for implementing (among others) Recommendation 6-2 of R&R: “Academic institutions and institutions managing scientific work such as industry and the national laboratories should include training in the proper use of statistical analysis and inference. Researchers who use statistical inference analyses should learn to use them properly.” R&R emphasized the reason for this recommendation: “While researchers want and need to use these tools and methods, their education and training have often not prepared them to do so.”

And that is exactly what education at all levels and fronts can help with, and what each of us can contribute, from being directly involved in an educational endeavor to just being a responsible citizen, disseminating reliable information and promoting effective, early, and unprejudiced education.

Acknowledgments & Disclosures

I am grateful to NASEM for their collaborations that made this special theme on R&R possible, and to Microsoft and the [MS Open Data Campaign](#) for partial support through the Harvard Data Science Initiative's Trust in Science project. I thank all the authors for their great contributions to *HDSR*, and for their comments and corrections on this editorial. I am continuously in deep debt to many *HDSR* board members for their on-going support, and especially to Radu Criaeu, Mark Glickman, and Robin Gong for keeping me editorial readable. Of course, all remaining defects are mine, and I sincerely hope that they are neither reproducible nor replicable. I also thank Mark for the inviting picture below, and I'd leave to readers to figure out with which parts of the data it should be paired.

Appendix

Reproducible and Replicable Happiness

As I mentioned earlier, it's impossible for me to end this editorial without reporting how reproducible and replicable was the “[Algorithmic Recipe](#)” given at the end of my last editorial. Here are the data I have, and I will leave it to all interested readers to draw their own conclusions, including how reliable these data are.

I just made it! It was delicious and crunchy. I used the soy sauce and wasabi combo and also diced a green onion in there with a bit of grated ginger- really added a zing to it.

What fun! It's a fantastic column!

I made the okra - it came out great and crunchy! I boiled the okra for $x=80$ seconds, and then chilled the okra in the refrigerator for $y=20$ minutes. I made a dipping sauce out of soy sauce, toasted sesame oil and wasabi. I cut up the okra before eating. I'll likely have a bunch tomorrow since I have dipping sauce leftover.

Thanks again for sharing the recipe!

A bit late to report, but, I made it Thursday ($X = 1.6$, $Y = 22.5$) and I was honestly pretty blown away. I love the taste of okra but can't stand the sliminess, and this really worked to eliminate the slime. One

small tweak I made without thinking that certainly didn't negatively affect the results: I threw a good amount of salt into the water.

Anyway, thank you for sharing this, and congrats on another great issue launch!

For what it's worth, I also (instinctively) boiled the okra in salted water. With $X=1.33$ and $Y=20.0$, I had a miniscule amount of slime, but nothing that would stop me from making this snack again.

I write this as we decided to try out your recipe from HDSR and ENJOYING it! OMG; its amazing! I am not nearly a connoisseur on the wine side, but Prosecco seems just fine too....THANK YOU. I will be curious to see how many write to you on having tried this! And you should talk about it in your next editorial!

Love the clarity and conciseness of your algorithmic recipe! As a former physicist, I think it might be worth trying faster chilling in ice water, because of the larger heat transfer and heat capacity of water compared to air. Curious if that results in any noticeable change in crunchiness.

The okra recipe is hilarious! Sadly, it doesn't seem like a jointly convex problem in X and Y , but luckily, that calls for a lot of experimentation and potential for surprising results, kind of like deep learning.

©2020 Xiao-Li Meng. This editorial is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the editorial.

Footnotes

1. This article is handled entirely independently by a co-editor who is otherwise not involved in editing this special theme on R&R, following the anonymous review policy of HDSR. All communications with the authors are done through me and my editorial assistant. [↵](#)

2. Because this analysis is based on a method I developed and I served in an advisory role to the authors, their submission's peer review process was blinded to me, and conducted by Liberty Vittert and Ryan Enos, who made all the editorial decisions. [↩](#)