We Should Do More Direct Replications in Science

Stuart Buck

***Column Editor's Note**:* Stuart Buck, in his contribution to this Reinforcing Reproducibility and Replicability column, ponders the value of direct replications and posits that more should be done, by funders, to support such direct replications. He argues that the value lies in part in letting such replication attempts tease out and test the documented steps, procedures, and mechanisms, removing the tacit or implicit knowledge that sometimes is present. Call this an audit, a verification, a test. In earlier contributions to this column, Butler (2023), Peer (2024), and Pérignon (2024) have shown that there is demand for such audits by researchers and institutions. Buck points out the role that funders have in supporting such efforts, because they, too, should care, if they want to achieve their goals of advancing science.

Keywords: reproducibility, replication, science, science policy, science funding, data analysis

---

Overview

Despite an arguable reproducibility crisis in many scientific fields, some have questioned the value of direct replications of prior studies. Their reasoning is that direct replications add little to our knowledge, and that we should focus on performing new studies. I argue, to the contrary, that direct replications are essential to scientific progress. Without direct replication, we have much less ability to know which prior scientific findings are actually worth trying to extend. As well, only direct replication can help us figure out puzzling anomalies about which contextual factors are important to a given scientific result.

Questioning the Value of Direct Replication

As we have seen over the past several years, there are problems with replicating the academic literature in many fields. The Reproducibility Project in Psychology found that only around 40% (give or take) of psychology experiments in top journals could truly be replicated (Open Science Collaboration, 2015). The Reproducibility Project in Cancer Biology similarly looked at studies from top journals in that field, and found that the replication effect was only about 15% as big as the original effect—for example, if an original study found that giving cancerous mice a particular drug made them live 20 days long, a typical replication experiment found that they lived 3 days longer (Center for Open Science, n.d.). Many pharmaceutical companies have said that they can barely replicate the academic literature, despite the fact that they have a huge incentive to carry forward

successful experiments into further drug development (Begley & Ellis, 2012; Prinz et al., 2011).

Due to these results and many others, a proposal I have made (Buck, 2022) is that science funders such as the National Institutes of Health (NIH) and the National Science Foundation (NSF)—which will spend nearly $60 billion this year, collectively—should dedicate at least 1/1,000th of their budgets to doing more replication studies. Even $50 million a year would be transformative, and would ensure that we can have higher confidence in which results are reliable and worth carrying forward into future work.

Oddly enough, not everyone agrees that directly replicating studies is a high-value activity. Indeed, when I was at a National Academies workshop recently, someone fairly high up at NIH told me that they were not in favor of funding direct replications. (It was a personal conversation, so I will not name the individual in question.)

The gist of this person's view was that we do not learn very much from trying to directly replicate experiments. After all, no experiment is ever going to be perfect, and we will find some discrepancies, but who cares? What really matters is whether the finding is robust in different contexts, so instead of funding exact replications, we should just fund new work that extends a prior finding in a new direction.

This NIH official is not the only one who is skeptical of the value of replication. Back when the Reproducibility Project in Psychology was finishing up in 2014, Jason Mitchell (2014) at Harvard famously wrote a short piece called "On the Evidentiary Emptiness of Failed Replications."

Mitchell's major claim is that it can be very hard to elicit a positive effect, and there are many more ways to mess up than to get things right. Moreover, there is a ton of tacit and unwritten knowledge in the fields of psychology and neuroscience (and, one presumes, other fields as well). By analogy, he says, if you take a recipe and follow it to the letter, but you do not actually know what 'medium heat' means or how to thinly slice an onion, you might not get the same results as an expert cook. But that does not mean the recipe is wrong, it just means that you do not have enough tacit knowledge and skill. Thus, he suggests, unless the replicators do everything perfectly, a 'failed replication' is uninformative to the readers.

We Need More Direct Replications

I would contend that *direct* replication of experiments in psychology, medicine, biology, economics, and many other fields, is highly useful and often essential to making progress. This is true for several reasons.

First, by doing direct replications (or at least *trying* to do so), at a minimum you learn how good a field is at disclosing its methods such that anyone else would be able to build upon a prior study.

With the Reproducibility Project in Cancer Biology (caveat: I funded that project while in philanthropy), we saw that literally zero percent of the time was it even possible to ***try*** to replicate a study (Errington et al., 2021). This was not because of tacit knowledge or because the original experimenters had some highly nuanced skill that the replicators lacked. Instead, it was because of obvious steps in the study that had to have happened, but that had not been documented very well at all.

For one example, "many original papers failed to report key descriptive and inferential statistics: the data needed to compute effect sizes and conduct power analyses was publicly accessible for just 4 of 193 experiments. Moreover, despite contacting the authors of the original papers, we were unable to obtain these data for 68% of the experiments" (Errington et al., 2021). In other words, they could not even figure out the magnitude of the effect they were supposed to be replicating. This is utterly basic information that ought to be included in any study.

Perhaps worse, "none of the 193 experiments were described in sufficient detail in the original paper" (Errington et al., 2021). In every single case, the team had to reach out to the original lab, which often was uncooperative or claimed not to recall what had actually happened in the study. For the 41% of the time that the original lab *was* cooperative, the answer was always that the replication team would need more materials and reagents (Errington et al., 2021).

That is why the entire project took longer, cost more, and completed fewer experiments than the project investigators had originally proposed when I funded this work while at the Laura and John Arnold Foundation. The quality of the literature was so low that it was impossible for anyone to fathom just how much effort and expense it would take even to *try* to replicate studies.

Clearly, the scientific literature can do better than this. All the top scientific journals should commit to publishing a truly *comprehensive* description of methods for every relevant study (including video as much as possible), so that others can more readily understand exactly how studies were conducted.

Second, if a study *is* successfully replicated, then we learn that we can have more confidence in that line of work. With the possibility of significant irreproducibility (see above), and even outright fraud on occasion, it is good to know what to trust. For example, last year *Science* published a lengthy story detailing how a prominent Alzheimer's study (Lesné et al., 2006) was likely fraudulent. To quote from the *Science* article (Piller, 2022):

> The authors "appeared to have composed figures by piecing together parts of photos from different experiments," says Elisabeth Bik, a molecular biologist and well-known forensic image consultant. "The obtained experimental results might not have been the desired results, and that data might have been changed to … better fit a hypothesis."

Nobel Laureate Thomas Sudhof (a neuroscientist at Stanford) told *Science* that the "immediate, obvious damage is wasted NIH funding and wasted thinking in the field because people are using these results as a starting point for their own experiments"

(Piller, 2022). A systematic replication project in Alzheimer's might have turned up that fact long before now. Researchers in that field would have had a better idea as to which studies to trust, and where to try to explore further.

Third, there is always the possibility that a study cannot be replicated very well or at all. Let us take a specific example from the Reproducibility Project in Cancer Biology. The results were that "[r]eplication effect sizes were 85% smaller on average than the original findings. 46% of effects replicated successfully on more criteria than they failed. Original positive results were half as likely to replicate successfully (40%) than original null results (80%)" (Center for Open Science, n.d.).

Contrary to Harvard's Jason Mitchell and the NIH official who spoke with me, I do think we can learn a lot from 'failed' replications. As an initial matter, it is possible that the replication team is incompetent, or does not have enough tacit knowledge, or made a simple mistake somewhere. But it does not seem likely to be true in all cases. Indeed, the replicators might often be *more* skilled than the original investigators. And when we know that so many pharma companies cannot replicate more than one-third of the academic literature—despite highly qualified teams who have every incentive to come up with a successful replication so that the program can move forward—it seems like we have bigger problems than 'replicator incompetence.'

Another possibility is that the original study cannot be fully trusted for any number of reasons. Perhaps there was improper randomization, improper treatment of outliers, questionable use of statistics, publication bias, *p*-hacking, outright fraud, or just a fluke. To be sure, we do not *know* any of that just because of one failed replication. But we do have a reason to suspect that the original result is not the full truth. Indeed, just due to publication bias alone, we might not want to trust the original publication even if everything had been done correctly.

A third possibility is that the original study and the replication are *both* correct, but there is some subtle difference in context, population, and so on, that explains the difference in results. Consider the classic paper of Hines et al. (2014), in which two labs on opposite coasts of the United States tried to work together on an experiment characterizing breast cancer cells, but found themselves stymied for a year or so during which their results were inconsistent. By traveling to each other's labs, they finally figured out that, unbeknownst to anyone in the field, the rate of stirring a tissue sample could change the ultimate results. They would never have known that the rate of stirring was important unless they had been trying to exactly duplicate each other's results.

It seems hugely important to know which seemingly insignificant factors can make a difference. Otherwise, someone trying to extend a prior study might easily attribute a change in results to the wrong thing!

Thus, we have many reasons to think that direct replication of a scientific study (or of a company's data analysis) is important. A direct replication can expose flaws in how the original study was reported, can expose faulty practices (or even fraud), can help us know how to extend a prior study to new areas, and at a minimum can help us know which results are more robust and trustworthy.

In short, I believe there is a clear case for devoting a small share (perhaps one tenth of one percent) of the federal government's research funding to direct replication. Future research would be more reliable, productive, and innovative, and will lead to more pharmaceutical cures than a system that puts 100% of the research dollars toward new research, while ignoring direct replication.

Disclosure Statement

The author works for an organization (Good Science Project) dedicated to improving science, but has no financial conflicts of interest as to any topic discussed here.

References

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature, 483*(7391), 531–533. https://doi.org/10.1038/483531a

Buck, S. (2022, October 4). Why we need more quality control in science funding. *Good Science Newsletter*. https://goodscience.substack.com/p/why-we-need-more-quality-control

Center for Open Science. (n.d.). *Reproducibility Project: Cancer Biology*. https://www.cos.io/rpcb

Errington, T. M., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Reproducibility in cancer biology: Challenges for assessing replicability in preclinical cancer biology. *eLife*, *10*, Article e67995. https://doi.org/10.7554/eLife.67995

Hines, W. C., Su, Y., Kuhn, I., Polyak, K., & Bissell, M. J. (2014). Sorting out the FACS: A devil in the details. *Cell Reports, 6*(5), P779–781. https://doi.org/10.1016/j.celrep.2014.02.021

Lesné, S., Koh, M. T., Kotilinek, L., Kayed, R., Glabe, C. G., Yang, A., Gallagher, M., & Ashe, K. H. (2006). A specific amyloid-β protein assembly in the brain impairs memory. *Nature, 440*(7082), 352–357. https://doi.org/10.1038/nature04533

Mitchell, J. (2014, July 1). *On the evidentiary emptiness of failed replications*. https://bpb-us-e1.wpmucdn.com/websites.harvard.edu/dist/2/77/files/2022/05/Mitchell_failed_science_2014.pdf

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 1–8. http://doi.org/10.1126/science.aac4716

Piller, C. (2022). Blots on a field? *Science, 377*(6604), 358–363. http://doi.org/10.1126/science.add9993

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews: Drug Discovery, 10*(9), Article 712. https://doi.org/10.1038/nrd3439-c1
