Neuron
NeuroView

The Costs of Reproducibility

Russell A. Poldrack[1,*]
[1]Department of Psychology, Stanford University, Stanford, CA, USA
*Correspondence: poldrack@stanford.edu
https://doi.org/10.1016/j.neuron.2018.11.030

Improving the reproducibility of neuroscience research is of great concern, especially to early-career researchers (ECRs). Here I outline the potential costs for ECRs in adopting practices to improve reproducibility. I highlight the ways in which ECRs can achieve their career goals while doing better science and the need for established researchers to support them in these efforts.

One of my favorite aspects of giving a colloquium talk at another institution is the lunch with early-career researchers (ECRs) that usually accompanies such a visit. In the last few years, the conversations during these lunches have invariably turned to reproducibility. This generation feels the ongoing reproducibility crisis acutely, and there is widespread enthusiasm for improving research practices among the current generation of ECRs. However, there is also a great deal of concern about the potential impact of adopting these best practices on their careers, given that career incentives are not yet aligned with these practices. My goal in this NeuroView is to address some of these questions—with a warning up front that many of the answers will be only partially satisfying at best. First, I will address how ECRs can manage to do reproducible science in the face of the resource constraints that many of them face. Second, I will address how ECRs can navigate the current incentive structures. Finally, I address how my fellow established researchers can help change the incentives to promote the work of ECRs who adopt more reproducible practices.

The Futility of Underpowered Science

One of the most common issues that comes up in these discussions is statistical power. It has become clear that low statistical power is a problem throughout neuroscience (Button et al., 2013). Running a weakly powered study is basically futile, since the likelihood of finding a positive result, even if an effect truly exists, is small. In fact, in clinical trials it's common to use what's called "futility analysis" to determine whether to discontinue a study that has no hope of finding a positive result. Futility wastes time and money, but an even worse effect of low power is that it reduces the likelihood that any positive findings are true (known as the positive predictive value). A thought experiment helps to explain this. Imagine running a study that has zero statistical power—that is, it has no ability to find a true positive result, even if one exists. For example, let's say that our detector is broken so that we are only measuring random noise. There will still be some number of false positives, which in standard null hypothesis testing is controlled at some low rate (usually 0.05). However, these are the only positive results that will occur, meaning that all the positive findings are false. As power increases, the proportion of true positive to false positive results increases, such that greater statistical power gives one greater confidence in the positive findings of a study.

The primary determinant of statistical power that is under our control is sample size, and it has become increasingly common for studies to increase sample sizes in order to address concerns about power. However, this can put ECRs in a bind, especially those using expensive techniques and/or training in labs without sufficient resources to collect well-powered samples. They need to complete a dissertation and desperately want to do reproducible work, but often simply don't have the resources to collect a sufficiently large sample—what's a student to do? I think the answer is not to collect an underpowered sample and label it a "pilot study." Rather, there are several strategies that students can take to succeed in such an environment.

Pivot. In a reflection on her work that characterized face responses in the human brain, Nancy Kanwisher (2017) discussed how she was driven to pivot from her initial interest in object perception to a more focused (but, to her, less interesting) study of face perception:

> I had never worked on face perception because I considered it to be a special case, less important than the general case of object perception. But I needed to stop messing around and discover something, so I cultivated an interest in faces. To paraphrase Stephen Stills, if you can't answer the question you love, love the question you can.

If you can't perform a sufficient study of the thing you are most interested in, then perhaps try to find a related question where you can make progress. The points below provide several examples of such pivots.

Collaborate. An instructive example of solving the power problem through collaboration comes from the field of genetics. Two decades ago it was common to see genetic association studies with small samples, often demonstrating implausibly large effects of candidate genetic variants. It has since been realized that most of these findings were probably false (Flint and Munafò, 2013). In place of these small studies, the field has moved to genome-wide association studies, which require very large sample sizes in order to be sufficiently powered to find associations after stringent statistical thresholds are applied. Few labs can afford to collect data from tens of thousands of individuals, but rather than simply accept underpowered studies, researchers in that field have instead formed large consortia that together

have generated datasets large enough to powerfully test the hypotheses of interest and have demonstrated replicable associations. Such collaborations come with challenges—for example, how does a researcher get credit for their work when they are listed among more than 100 other authors, as often happens on these papers? However, for some questions where many labs are collecting data, such large-scale collaborations may be the best solution to the power problem. A side effect of such a collaboration is that it gets one tied in with the relevant research community, which can have its own benefits in terms of career enhancement.

Use shared data. The amount of openly available data in neuroscience is staggering, especially in human neuroimaging, where data sharing has become commonplace, and shared data have been used in many highly impactful publications (Milham et al., 2018). There are certainly many questions that can't be asked using shared data, but if your question (or a related one, as noted above) can be answered using openly available data, that can be a very fruitful alternative to collecting your own data. There are many success stories in human neuroscience of researchers who have succeeded by bringing interesting new ideas to bear on openly available data. For example, one recent postdoc in my lab was able to publish a set of papers in journals including *Neuron*, *Nature Neuroscience*, and *PNAS* based completely on shared data collected by other groups. If you can't find relevant data, then consider contacting another lab and asking to obtain their published data for further analysis. This will sometimes fail, as many researchers remain unwilling to share data (Wicherts et al., 2011), but when it works it can provide you with free data, as well as a chance to interact directly with other researchers who are focused on the same question.

Embrace theory. Data are of course necessary for good science, but they are far from sufficient. A striking example of this comes from the final paragraph of Kasthuri et al. (2015), who state the following after presenting a mind-blowing demonstration of the structural complexity of a tiny volume of cortical tissue:

> Although technologies, such as the ones described in this paper, seek to provide a more complete description of the complexity of a system, they do not necessarily make understanding the system any easier. Rather, this work challenges the notion that the only thing that stands in the way of fundamental mechanistic insights is lack of data.

In the context of the BRAIN Initiative, which has funded the development of incredible new tools for data acquisition, it is increasingly clear that deeper theory and more powerful computational methods will soon become the limiting factors in the progress of neuroscience. Thus, another move for trainees without the ability to collect sufficient data is to pivot to theory and/or computation. In nearly every domain, there are open questions about how to model and analyze data, and computational neuroscientists are in great demand right now, not least because so many of them are being hired by tech firms like Google and Facebook. A bonus of this kind of training is that the analytic skills that you learn will still be very useful if you later return to hands-on experimentation, as well as being highly transferable outside of academia.

The Incentives Are Not Yet Aligned

Many of the technical methods proposed by reproducibility advocates, such as version control and automated analysis pipelines, have the potential to greatly improve a researcher's efficiency. However, in my discussions with ECRs, I also try not to sugarcoat the fact that some of the remedies we advocate are likely to make them less competitive on the job market in the short term. Larger sample sizes and replication of each study means fewer studies, which means fewer papers on their CVs. Pre-registration hamstrings one's ability to tweak the analyses after seeing the data, which will greatly improve reproducibility but is almost certain to make it harder to generate splashy findings that can land papers in high-profile journals. Given this, why should an ECR even bother? I have two answers. The first is that once the field catches up (and I am certain it will), then you will be viewed as one of the pioneers. The second, deeper answer is embedded in a quote from the physicist Richard Feynman in his famous 1974 Caltech commencement address (http://calteches.library.caltech.edu/51/2/CargoCult.htm) on the integrity that distinguishes one as a scientist:

> It's a kind of scientific integrity, a principle of scientific thought that corresponds to a kind of utter honesty—a kind of leaning over backwards. For example, if you're doing an experiment, you should report everything that you think might make it invalid—not only what you think is right about it: other causes that could possibly explain your results; and things you thought of that you've eliminated by some other experiment, and how they worked—to make sure the other fellow can tell they have been eliminated. Details that could throw doubt on your interpretation must be given, if you know them. You must do the best you can—if you know anything at all wrong, or possibly wrong—to explain it.

Unfortunately, the current incentives toward a large quantity of high-impact publications cut directly against this kind of integrity. I think that most of us became scientists rather than bankers or fortune-tellers because we wanted to be able to experience the joy of generating new knowledge about the world, of spending our days asking interesting questions and obtaining answers in a way that we can trust. To allow the goal of reliable knowledge to be overtaken by careerism is to do what Feynman called "cargo-cult science," putting on an elaborate show in place of the true knowledge-building activity of science.

Of course, it's easy for me to say this. I have a permanent job at an elite institution, and high-profile journals like *Neuron* invite me to write pieces like this, not to mention many other privileges including gender, race, and national origin. I'm not a grad student or postdoc with loads of college debt, worried about whether I will get a faculty job just so I can provide myself and my family with a stable life. Feynman, too, realized the privilege of

his position and the challenge of maintaining integrity in the face of real life:

> So I wish to you—I have no more time, so I have just one wish for you—the good luck to be somewhere where you are free to maintain the kind of integrity I have described, and where you do not feel forced by a need to maintain your position in the organization, or financial support, or so on, to lose your integrity. May you have that freedom.

A particular concern that is occasionally raised is "ladder-pulling"—i.e., that it is disingenuous for researchers like myself to demand improved practices when our careers were built on the kind of small-sample studies that we now tell researchers they shouldn't do. I am sympathetic to this argument and have tried to be open about what I see as shortcomings in our previous methods (Poldrack, 2018). But I would also point out that science is always a moving target, and that it's inevitable that some of our practices will be found to be lacking as science moves forward. It's our responsibility to do the best we can given what we know right now and to embrace the best possible methods. And it's particularly the responsibility of those of us who built our careers on methods we now realize to be outdated to advocate for those ECRs who wish to do better—as I outline further below.

When ECRs ask why they should have hope that their reproducible practices will pay off, I can only say that I am hopeful that the world is changing. Researchers of my generation, who are responsible for many hiring and tenure decisions at this point, are becoming increasingly aware of the reproducibility problem, and this is starting to flavor our decisions. However, it's also worth realizing how challenging this kind of change is and admitting that there are no guarantees.

Making Your Efforts Known

There are several things that ECRs can do to signal their efforts toward reproducibility. First, they should consider adding a section to their CV that highlights their relevant activities, including shared data or code, pre-registrations, and any other relevant activities. Second, get involved in groups that are committed to open science and reproducibility; if there isn't a local group, then consider starting one. Within the neuroimaging field, many trainees have entered this world through hackathons, which provide researchers at any level of programming skill with the opportunity to get involved in software development projects and, more importantly, engages them in a community of like-minded individuals. Engaging in these communities can also provide support and encouragement for early-career researchers who want to engage in reproducible and open practices but who may face resistance from supervisors who are skeptical or even openly hostile to these changes.

Getting involved in social media is another way to make yourself known as a member of the community. For me, Twitter has become an invaluable way to interact virtually with many junior scientists who I never would have known otherwise. One should not discount the power of this kind of familiarity in processes like job searches—search committees must often dig through hundreds of applications in order to come up with a short list for deeper review, and anything you can do to make your name pop out of this initial pile is likely to increase your chances. Having a large number of Twitter followers will not get you a job on its own, but when your name pops out to a member of the search committee, it will almost certainly help your chances of getting a second look that could help bump you onto the short list. In addition, social media discussions have led directly to collaborative papers, so getting involved in these discussions can also pay off on your CV.

Colleagues, Let's Admit that We Have a Problem

So far I've been speaking to trainees, but it seems equally important to highlight the things that my peers (i.e., faculty members making decisions about hiring and tenure) can do to fix the reproducibility problem. First, we must acknowledge that we have a problem, and that all of our careers have benefitted from the research practices that we now realize to be questionable. Although reproducibility problems in psychology have received the most publicity, the reproducibility problem appears to be widespread; for example, a 2016 poll (Baker, 2016) found that more than 80% of chemists had failed to reproduce results published by another group at least once, with biologists close behind at more than 70%. As the gatekeepers of the academy, it's our responsibility to fix this problem.

Most critically: when our trainees express their wishes to improve research practices, we must give them our full support. As much as we would like to think that everyone would agree with this, the "Bullied into Bad Science" movement (http://bulliedintobadscience.org/) shows that ECRs regularly experience opposition from supervisors when they try to institute these practices.

We should also focus less on the results and more on the questions and methods. At present, it is exceedingly difficult to publish a negative result, even if the question is interesting and the methods are top notch, and if it is published, it is likely to be in a journal of low impact. This focus on positive results is arguably one of the central drivers of the reproducibility crisis. A bit of reflection makes clear that if every study we are doing has a positive result, then either we are doing uninteresting science with obvious outcomes, or our methods are somehow skewed such that we are finding positive results even when there is no true effect (and throwing everything else in the garbage can). In particular, requiring a graduate student to find a positive result in order to graduate (which is seemingly still common) provides an exceedingly strong incentive to engage in questionable research practices and should be abandoned. Training should focus on learning how to ask important questions using robust methods, not on how the answers to those questions come out.

Finally, we need to lead by example, aspiring to demonstrate for our trainees the kind of integrity that Feynman spoke of as the *sine qua non* of science. We need to make clear that getting it right is more important than getting it published in a high-profile journal and that the answer to an important question is important regardless of whether it's positive or negative. Many of us are fortunate enough to be in that position of freedom, and it's our responsibility to do everything we

can to advance the careers of ECRs who are focused on scientific integrity.

ACKNOWLEDGMENTS

Thanks to Rotem Botvinik-Nezer, Danilo Bzdok, Mike Frank, Chris Gorgolewski, Konrad Körding, Dan Lurie, David Mehler, and Jessey Wright for helpful comments on a draft of this article, as well as to the Twitter community for helpful discussion.

REFERENCES

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature *533*, 452–454.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., and Munafò, M.R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. *14*, 365–376.

Flint, J., and Munafò, M.R. (2013). Candidate and non-candidate genes in behavior genetics. Curr. Opin. Neurobiol. *23*, 57–61.

Kanwisher, N. (2017). The Quest for the FFA and Where It Led. J. Neurosci. *37*, 1056–1061.

Kasthuri, N., Hayworth, K.J., Berger, D.R., Schalek, R.L., Conchello, J.A., Knowles-Barley, S., Lee, D., Vázquez-Reina, A., Kaynig, V., Jones, T.R., et al. (2015). Saturated Reconstruction of a Volume of Neocortex. Cell *162*, 648–661.

Milham, M.P., Craddock, R.C., Son, J.J., Fleischmann, M., Clucas, J., Xu, H., Koo, B., Krishnakumar, A., Biswal, B.B., Castellanos, F.X., et al. (2018). Assessment of the impact of shared brain imaging data on the scientific literature. Nat. Commun. *9*, 2818.

Poldrack, R.A. (2018). The New Mind Readers: What Neuroimaging Can and Cannot Reveal about Our Thoughts (Princeton, NJ: Princeton University Press).

Wicherts, J.M., Bakker, M., and Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. PLoS ONE *6*, e26828.