The replication challenge: Is brain imaging next?

11

David Mehler

The concept of replication is simple. Do you remember high-school chemistry class—mixing compounds to create explosions? Whether you knew it or not, you were engaged in an attempt to replicate the findings of previous scientists. Unlike these pedagogic experiments, researchers now increasingly attempt to confirm the veracity of findings with carefully crafted and meticulously executed replication studies.

In psychology, for example, a collaboration of unprecedented size set out to replicate 100 highly influential behavioral studies [1]. They successfully replicated 39. As for the remaining 61 studies, we can't be certain whether the original results represent true effects. While these problems of replication have likely existed for decades, only recently have scientists become highly aware of them and realized that—ironically—one of the most replicable findings across the life sciences is the difficulty of replication itself! In the field of brain imaging, which has considerably fewer replication studies than psychology, researchers have begun to understand why our field may suffer from a similar syndrome. This chapter discusses the forces behind low replicability, including publication bias and researcher bias, and then highlights how adopting an incentive system that encourages high-quality research practices can help overcome these issues.

While some neuroimaging findings seem robust enough to forego formal replication attempts, others could use a second look. For example, motor areas will reliably activate during movement and the amygdala (an area involved in processing fear) will surely respond to fearful stimuli. For several reasons, however, we can't be so certain that more complex analyses such as brain−behavior relationships and contrasts between participants will replicate [2] (e.g., patients compared to healthy

Casting Light on the Dark Side of Brain Imaging. DOI: https://doi.org/10.1016/B978-0-12-816179-1.00010-4

controls). First, head motion, respiration, and heartbeats all contaminate brain recordings and require meticulous removal before analyzing the data [3] and performing statistical analyses [4] (see Chapter 9). Researchers may mistake this noise for a neural signal, especially in "resting state" studies which observe spontaneous brain activity (see Chapter 23). Second, publication bias, researcher bias, and low statistical power all decrease the probability that positive findings represent true effects [5]. *Meta-research*—an entire field in its own right—studies these phenomena and their implications.

Unlike most media reports we hear these days, the press on brain imaging tends to be quite positive. This trend may emerge from *publication bias*—where positive findings are much more likely to reach publication than null results [6]. And yet, null results aren't necessarily the bad news that many people cut them out to be; they may be as informative as positive results. They can tell researchers that the phenomenon they are testing for may not exist after all and that their time can be better spent elsewhere. There are several reasons for publication bias. For instance, researchers may not submit negative findings for publication because the results contradict their prior beliefs and journals tend to decline the publication of null findings more often than positive results [7]—both examples of what scientists call the *file drawer effect*. Thus the published literature may merely represent the "tip of the iceberg" [8]. A massive amount of research may never surface, and in turn, never inform future experiments. How we report scientific findings appears to matter more than previously thought.

Another factor, *researcher bias*, also reduces replicability. This term encompasses various questionable research practices, including selectively presenting results that fit a preferred storyline and omitting information that would allow others to replicate an experiment. Investigators often adopt questionable research practices unintentionally, but at times they may also intentionally modify their analyses to push statistical results beyond the widely accepted line that scientists use to define findings as significant (so called *p-hacking*) [9]. One common example is when researchers perform many statistical tests on one dataset but fail to account for the fact that running more tests means that there will be a higher chance of obtaining spurious results. Every test is associated with a probability for a false positive, also called the *error rate*. In other words, test results may suggest that there is an effect although in reality no effect exists—that we are chasing ghosts. Statistical error rates increase with the number of tests conducted, so adjusting for multiple tests becomes especially important in brain imaging experiments, where researchers often perform thousands of tests on one dataset. Neuroimagers are still working out how to best adjust for multiple tests within complex analyses [4].

Clear reporting of methods is particularly important in brain imaging research. Take a guess at how many ways we can analyze data from a single brain scan. Theoretically countless, practically at least 69,000 different ways! [10]. Functional brain images aren't photographs. They are statistical maps resulting from complex digital image manipulation and analyses. In fact, brain imaging data usually require between 6 and 10 steps of general data preparation and analysis. Researchers can perform each of these steps in a variety of ways, regardless of whether the data

comes from functional magnetic resonance imaging (fMRI), electroencephalogram, magnetoencephalogram, or other imaging techniques. Different choices in data processing and analysis can lead to widely divergent results: small variations can quickly sum to form large discrepancies [10,11]. In some cases, researchers may run many variations of an analysis, but only report results that support their hypotheses. This practice can lead to biased publications that overestimate true effects [12]. The third major cause of poor replicability—low statistical power—is addressed in detail in the following chapter.

Now that we have covered the forces driving the publication of false positives, let's focus on how to remedy the problem. Current reward structures in academia—including job promotions and funding schemes—primarily incentivize publishing articles in esteemed academic journals. These high impact journals are much more interested in novel and glossy findings compared to replication efforts. Hence, the reward structure and publication system partly reinforce the status quo: scientists are incentivized to focus on positive findings and work with underpowered studies that consume fewer resources [13]. Further, they are encouraged to investigate novel effects, rather than retest old findings. In fact, such incentives may be one reason why scientists rarely design costly neuroimaging experiments with replication in mind. Overall, current reward structures largely fail to promote high-quality science.

What can neuroscientists learn from the replication challenge to make their work more replicable? Beyond more rigorous training in experimental design and statistics, one growing research practice may be game-changing—and that is the preregistration of the methods, analyses, and aims of an experiment before data collection even begins. This documentation is often stored on an openly accessible platform such as *clinicaltrials.gov* or the open science framework (OSF). Another option is to submit a *registered-report* where peer-reviewers assess the rationale for the study design, its proposed methods, and the planned statistical power before data are collected [14]. As long as authors follow their protocol, accepted registered reports are guaranteed to be published irrespective of the final results. Researchers can also upload open access versions of their articles before they reach publication, known as preprints. This procedure can allow for informal and transparent feedback, and hence an open discussion of the research. Altogether, preregistration, registered reports, and preprints help address the three main issues underlying the replication challenge: publication bias, researcher bias, and low statistical power.

In addition to collecting new data to *replicate* findings, scientists should also be able to reperform reported analysis on an existing dataset to *reproduce* results [11]. While replication studies are still relatively rare in neuroimaging (partly because imaging experiments are costly, but also because there are more incentives to study new research questions than probe previous ones), journals have recently published a number of insightful reproduction studies that used openly available data sets. Various factors influence how well results can be reproduced. These include the quality of documentation, potential errors in the original analyses, the robustness of the statistical effect [15], and even the software package used [16]. The more steps analyses comprise, the more likely errors can occur and sum up. Notably, the

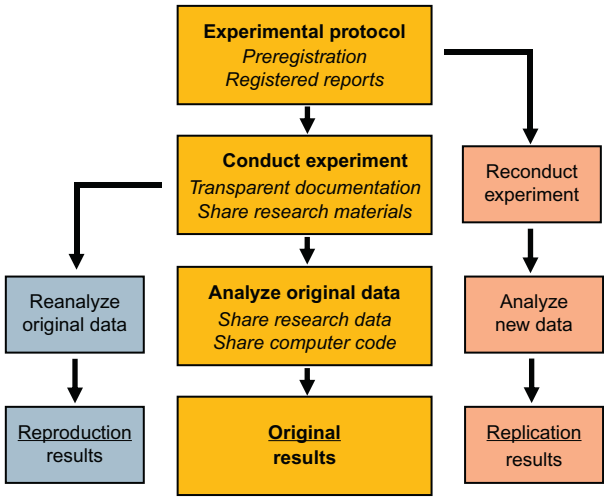

Figure 11.1 Depicts an original experiment (gold), a reproduction attempt (gray), and a replication attempt (orange). Whereas reproductions re-analyze the original data, replications re-conduct the original experimental protocol and analyze new data. Steps that can be taken to increase subsequent reproduction and replication success are listed in italics.

prestige of a journal falls short as a guarantor for the reliability of its findings. On the contrary, journal prestige may be associated with below-average reproducibility [17]. Only transparent reporting of methods, analyses, code, and data makes research more reproducible and hence represents best scientific practice [18].

Fortunately, the field of neuroimaging is witnessing rapid and promising efforts to improve the quality and robustness of its findings. Like research in psychology and genetics, neuroimagers are setting new standards with several open data projects (e.g., the human connectome project, open fMRI, and the Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA) consortium) [11]. Neuroscientists can use these open data sets to test parameters, validate analyses, and address new research questions [2,4,11]. They can also use these data to estimate the number of participants needed to render neuroimaging results replicable [19]. Altogether, these initiatives leave room for optimism. As the field moves toward more rigorous methodology, we can benefit from noting that replicating experiments alone can only tell us so much [20]. Instead, multiple lines of evidence from experiments that deliberately use different methods while addressing the same research question, known as *hypothesis generalizability*, are likely even more fruitful to advance science.

In summary, the replication challenge urges scientists across disciplines to overcome biases and maximize freely available access to their work in order to encourage independent researchers to test and verify published results. In doing so, researchers can benefit from openly sharing their hypotheses and experimental methods from the start. The field of neuroimaging in particular has started to address the replication challenge by developing best practice guidelines, software for reproducible science, and databases for preregistration and open science [11]. These developments will likely transform the way we image brains, and when combined with new technology, lead to promising breakthroughs.

Acknowledgments

The author acknowledges Dr. Chris Allen, Dr. Chris Chambers, Dr. Russell Poldrack, and Johannes Algermissen for insightful discussions on the topic, which have inspired parts of this chapter.

Additional readings

An in-depth review of questionable research practices and solutions to the replication crisis: Chambers C. The seven deadly sins of psychology: a manifesto for reforming the culture of scientific practice. Princeton university Press; 2017.

A landmark replication study in psychology, the Open Science Project: Collaboration OS. Estimating the reproducibility of psychological science. Science (80-) 2015;349(6251): aac4716.

A comprehensive review on reproducible neuroimaging: Poldrack RA, Baker CI, Durnez J, et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat Rev Neurosci 2017;18(2):115–26.