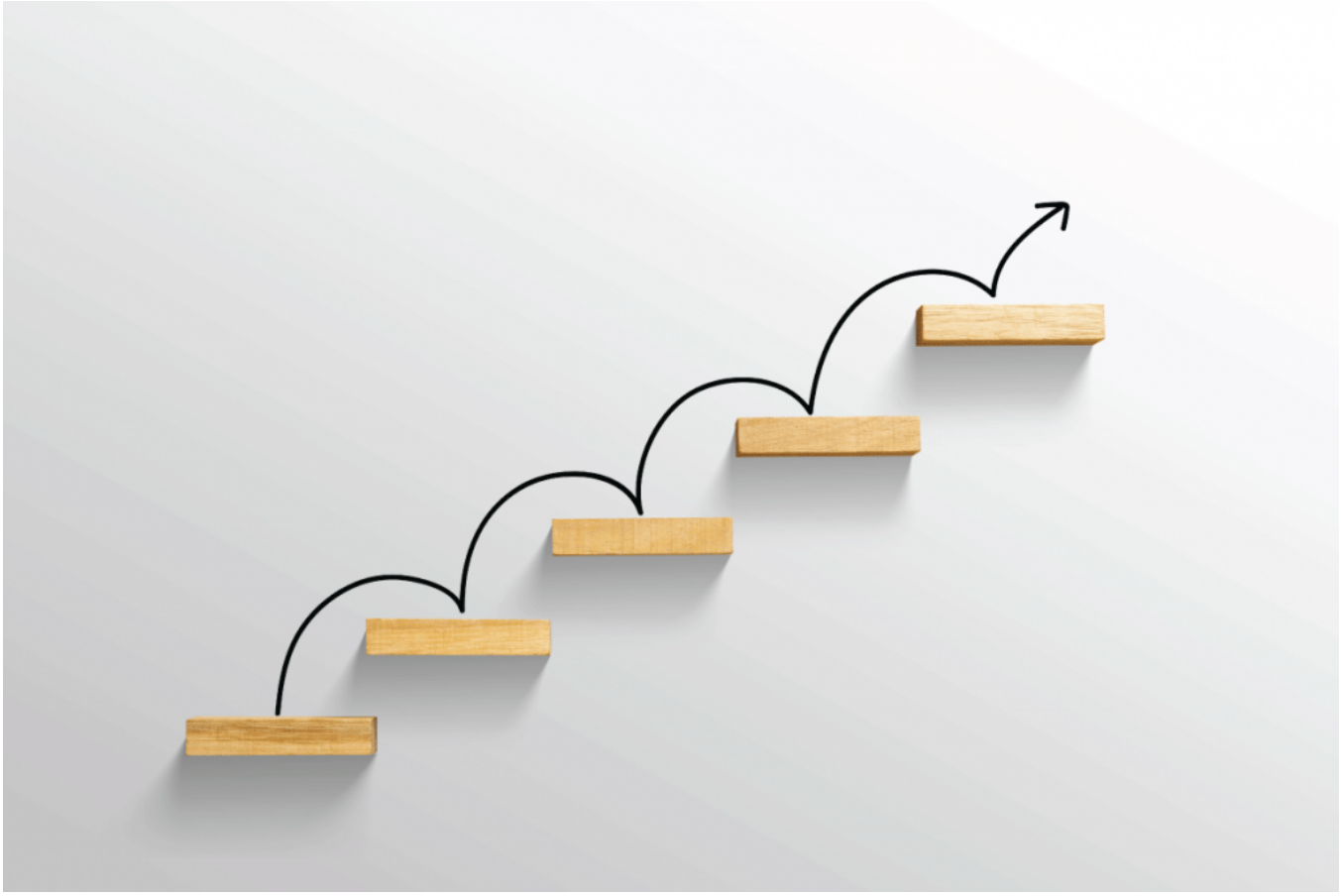
No Crisis but No Time for Complacency

August 22, 2019



The National Academies of Sciences, Engineering, and Medicine recently published a report titled Reproducibility and Replicability in Science. We both had the privilege of serving on the committee that issued the report, and this is a brief summary of how the committee came about and its main findings.

In response to concerns about replicability in many branches of science, Congress — via the National Science Foundation — directed the National Academies to conduct a study. The mandate was broad: to define reproducibility and replicability, assess what is known about how science is doing in these areas, review current attempts to improve reproducibility and replicability, and make recommendations for improving rigor and transparency in research — across all fields of science and engineering, not just psychological science.

A committee of 13 scientists was formed that, in addition to us, included geoscientists, medical researchers, natural scientists, engineers, computer scientists, historians of science, and statisticians. The committee met 12 times in a period of 16 months. This was not too difficult for Tim, who could hop on a train in Charlottesville and be in Washington in a couple of hours. It was more difficult for Wendy, who interspersed a sabbatical in Paris with flying back and forth to DC several times. Regardless, we both agree that it was a fascinating and enlightening experience to serve on the committee.

So, what did the committee conclude? Our job was first to define reproducibility and replicability. As you can imagine, definitions vary greatly across disciplines, and our consensus definitions were hammered out from a range of possibilities.

We defined *reproducibility* as computational reproducibility — obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis. *Replicability* was defined as obtaining consistent results across studies that were aimed at answering the same scientific question, each of which obtained its own data. In short, reproducing research involves using the original data and code, whereas replicating research involves new data collection and methods similar to those used in previous studies.

Once we defined our terms, what did the committee conclude about the state of reproducibility and replicability in science? This question is probably foremost in many people's minds, given the attention it has received, both in our field and in the national media. And, as anyone who has followed this debate knows, there is considerable disagreement about the answer. Some believe that our field faces severe problems, such as frequent use of lax methods, that threaten validity. Others feel that the extent of these problems has been exaggerated. Still other researchers note that rigorous research practices have been an important focus in psychological science and other scientific fields long before the current concerns with reproducibility and replicability.

The committee's answer was, in short, "No crisis, but no complacency." We saw no evidence of a crisis, largely because the evidence of nonreproducibility and nonreplicability across all science and engineering is incomplete and difficult to assess. At the same time, steps can be taken to improve in both areas.

The committee's specific conclusions and recommendations differed for reproducibility and replicability. One key difference involves the rates of reproducibility and replicability to which we should aspire. There is large agreement on the answer to this question for reproducibility: When a researcher transparently reports a study and makes available the underlying digital artifacts, such as data and code, the results should always be computationally reproducible. The committee made recommendations about how to achieve reproducibility, largely by improving transparency. For example, the committee proposed that, to help ensure the reproducibility of computational results, researchers should convey clear, specific, and complete information about any computational methods and data products that support their published results to enable other researchers to repeat the analysis.

The scientific ideal for replicability — in which researchers attempt to obtain consistent results by collecting new data, using similar methods — is more nuanced. For example, a key observation in the report, we believe, is that, "The goal of science is not, and ought not to be, for all results to be replicable" (p. 28), because there is a tension between replicability and discovery. (For an excellent discussion of this issue, see B. Wilson & Wixted, 2018, *Advances in Methods and Practices in Psychological Science, 1*, 186–197).

Similarly, the committee noted that nonreplicability can arise from a number of sources, some of which are potentially helpful to advancing scientific knowledge and others that are unhelpful.

Helpful Sources of Nonreplicability

Nonreplicability can be caused by limits in current scientific knowledge and technologies, as well as inherent but uncharacterized variabilities in the system being studied. When such nonreplicating results are investigated and resolved, it can lead to new insights, better characterization of uncertainties, and increased knowledge about the systems being studied and the methods used to study them.

Unhelpful Sources of Nonreplicability

Nonreplicability also may be due to foreseeable shortcomings in the design, conduct, and communication of a study. Whether arising from lack of understanding, perverse incentives, sloppiness, or bias, these unhelpful sources of nonreplicability reduce the efficiency of scientific progress.

One unhelpful source of nonreplicability is inappropriate statistical inference. Misuse of statistical testing often involves post hoc analysis of data already collected, making it seem as though statistically significant results provide evidence against the null hypothesis, when in fact they have a high probability of being false positives. Other inappropriate statistical practices include *p*-hacking — the practice of collecting, selecting, or analyzing data until a result of statistical significance is found — and "cherry picking," in which researchers may unconsciously or deliberately selectively report their data and results.

To minimize unhelpful sources of nonreplicability, we outlined initiatives and practices to improve research design and methodology, including training in the proper use of statistical analysis and inference, improved mentoring, repeating experiments before publication, conducting rigorous peer review, utilizing tools for checking analyses and results, and improving transparency in reporting.

Replicability and reproducibility are not the only ways to gain confidence in scientific results. Research synthesis and meta-analysis can help assess the reliability and validity of bodies of research. As you probably know, meta-analyses provide estimates of overall central tendencies (effect sizes or association magnitudes), along with estimates of the variance or uncertainty in those estimates. Meta-analytic tests for variation in effect sizes can suggest potential causes of nonreplicability in existing research — in individual studies that are outliers, in particular populations, or using certain methods. Of course, such analyses must take into account the possibility that published results are biased by selective reporting and, to the extent possible, estimate its effects.

To conclude on a personal note, it was fascinating to learn about the ways that different scientific disciplines attempt to establish reproducibility and replicability. We were more convinced than ever in the fundamental soundness of our field. Like other sciences, psychological science is producing a great deal of useful and reliable knowledge — replicable discoveries about human thought, emotion, and behavior. Increasingly, researchers and governments are using such knowledge to meet social needs and solve problems, such as improving educational outcomes and reducing government waste from ineffective programs. We strongly endorse the broad conclusion from our meetings: No crisis, but no time for complacency!