Opinion Page

Opinion: *Is Reproducibility a Crisis for Science?*

Bradley E. Alger
Professor Emeritus, Department of Physiology
University of Maryland School of Medicine
balgerlab@gmail.com

Many of us who follow developments in science have been alarmed to read in scientific journals and major national news media that science is suffering a "reproducibility crisis." Reproducibility, or replicability, is the property that allows scientific findings to be repeated. Reproducibility is a core value of science and the apparent widespread failure of prominent findings to be reproduced is seen as a serious threat to the integrity of the whole enterprise. The big fear is that irreproducible science is bad science. How can we accept the claims of climate scientists and vaccine developers if their science rests on shaky ground? It's a fair question, however reproducibility is a surprisingly subtle and complex topic, and judging the seriousness of the problem requires a good grasp of the details. In this essay, I'll examine reproducibility in its various forms, its significance, and its place in scientific thinking and reasoning. And I'll argue that, while true scientific claims will be reproducible, the pursuit of reproducibility *per se* is not always the best guide for conducting research. I'll consider challenges that bench scientists must grapple with when confronted with irreproducible results in their fields. The impact of either reproducibility or irreproducibility is often less than what we've been led to believe. Clarifying some practical issues should help lessen the anxiety about reproducibility and bolster our willingness to trust in science.

What is "reproducibility"?

To begin at the beginning: what do we mean by the "reproducibility" of scientific findings? I follow the American Society for Cell Biology in distinguishing four distinct senses in which the word is used, which I'll illustrate with a made-up example. Suppose you've done a study on how Drug X affects the ability of rats to learn to navigate a maze. Following best scientific practice, after the experiments you make your raw data, say the number of trials taken by each rat to master the maze, available on an accessible website. If I re-examine your data, subjecting it to the same or different statistical analyses, replotting it, etc. I am carrying out a test of *analytic* reproducibility. I don't do any new experiments, I am simply checking to see whether the conclusions that you've drawn from your data are accurate and valid.

Now suppose that I set out to repeat your exact experiment faithfully. I acquire rats, mazes, measuring instruments that are as like yours as possible and I redo your experimental procedures. This would be a test of *direct* reproducibility; it seeks to know whether your actual findings can be replicated. This is probably what most people think of when they hear about reproducibility.

Two other experimental approaches that are sometimes used in "reproducibility" studies differ markedly from analytic and direct reproducibility but are lumped together with them. This considerably muddies the waters surrounding the problem.

Imagine that I adopt most of your experimental conditions except one or two. Perhaps instead of rats, I elect to study mice because I hope eventually to take advantage of the greater opportunities for genetic analyses of behavioural phenomena that the mice afford. If the mouse results do not replicate the rat results, is this a failure of reproducibility? Obviously not: mice are not rats and differ from them in many ways. (Rats get up in the morning, grab their lunch pails, and get to work learning mazes. Mice are more free-spirited, less focused. After a bit of earnest maze-learning, they may suddenly go back and retrace a previously unrewarded path just because they're mice.) Maybe the drug simply affects mice differently than it does rats. Although there are myriad possible rational explanations for the divergent behavioural outcomes, follow-up studies in which original conditions are deliberately varied have been rated as failures of *systematic* reproducibility.

Finally, suppose that you've interpreted your results as meaning that the drug adversely affected the rats' ability to learn mazes because it caused them physiological stress. That's your hypothesis, and it would predict, for example, that the drug will alter the rats' stress hormone levels. If my follow-up study finds that rat blood levels of the appropriate hormones are unchanged, some writers would classify this as a failure of the *conceptual* reproducibility of your results. Your hypothesis was evidently false, and that's why I didn't get the results it predicted. Indeed, I might even have reproduced your empirical findings, while showing that your idea flunked the conceptual test.

Of the four supposed classes of "reproducibility," only the first two, *analytic* and *direct*, involve actual attempts to replicate prior observations. The latter two, *systematic* and *conceptual* reproducibility, plainly do not and should not be weighed in the reproducibility debate at all, if the goal is to assess the reliability of scientific findings. On the contrary, systematic and conceptual studies are explorations of inferences that are derived from original results; in fact, they expand knowledge. Ordinarily, they are tests of hypotheses that were explicitly or implicitly suggested by the earlier studies. Varying conditions or deducing and testing predictions is how scientists decide on the merits of their hypotheses. Hence, a failure of systematic or conceptual reproducibility is no cause for alarm; it is an illustration of the scientific method in action. When debating the reality of a "reproducibility crisis" we must be clear on what sort of experiments we're talking about and to zero in on only those that check for analytical or direct reproducibility. In the remainder of this essay, "reproducibility" refers to these two.

Unrecognised variability in experimental materials is a common cause of irreproducible results. Culprits include impurities or unwarranted differences in nominally the same chemical reagents, animal strains, in vitro "cell lines," antibodies, and

so forth. It is no wonder if original results obtained from experiments done on, e.g., isolated brain cells are not duplicated in a would-be follow-up study carried out on mislabeled skin cancer cells. While deplorable, errors of this kind are essentially quality-control problems: fixable with proper attention to detail, analysis, screening, and improvement in the production processes. We must be aware of them and can work to eliminate them without losing confidence in science as a whole and I won't consider them here. Likewise, I will not deal with scientific misconduct, cognitive biases, or issues related to the distorted reward system of science. These are thorny psychological and policy issues that deserve their own coverage and are outside the scope of this essay.

What defines "reproducibility" and how much irreproducibility must we tolerate?

What does it mean to fail to reproduce a study? Perfection is not an option in science; there is no reasonable chance that a second study will obtain the exact numerical results, say the identical mean and standard deviation, of a group of measurements, as did the first study. The rich variability of the world, which encompasses investigators and measuring instruments as well as things measured, precludes exact reproducibility. To sidestep the problem, scientists decide that if two results are "close enough" then for all intents and purposes they are the same. And they use statistics to define "close enough;" in the current context, whether a second study has reproduced the first. The main point here is that reproducibility is a statistically determined property, not something carved in stone. It follows as a corollary of the probabilistic framework that 100% reproducibility – every attempt succeeds – is unachievable. So how much irreproducibility is unavoidable; how much must we tolerate?

There are several ways of estimating the degree of reproducibility that we should expect. As an illustration, we'll review one that uses traditional concepts of *p-values* and *statistical significance*. To oversimplify greatly, scientists usually calculate the probability that a given experimental result would have arisen by chance alone given a few generally plausible assumptions about how random variability enters the picture. They use statistics to determine if the odds of the result's happening by chance. If the odds are small, they reject that idea and tentatively attribute the result to their experimental treatment. Similarly, since no two experimental groups will be absolutely identical, at some level of measurement precision they'll differ. Hence, scientists look for a difference that is *big enough* to persuade themselves and others that the apparent treatment effect is real. If the performance of the drug-treated rats is different enough, *significantly different*, from that of the untreated rats, experimenters conclude that the drug probably did something. How rare a chance event has to be before they make this leap is a matter of convention and judgement; it varies across scientific fields. Biologists are usually willing to consider a significant event as one that would occur by chance on only 1/20 of the trials or less (the probability, p-value, is ≤ 0.05); particle physicists are much more conservative, holding out for a probability of 1/3,500,000 ($p \leq 0.0000003$), before they'll acknowledge the discovery of a new physical entity (a standard the Higgs Boson famously met in 2012). The p-value is also called the significance level. No matter what its specific value, however, the p-value is merely an estimate of the probability that they're making a mistake in thinking they've discovered a real effect.

With p≤0.05 as a standard, even with correct methods and optimal conditions, there would be a 5% chance of wrongly concluding, e.g., that the drug affected the rats' maze-learning ability, and therefore these results are unlikely to be reproducible. The odds that another group would be victimised by the same bad luck is 1/20 x 1/20 = 1/400. In other words, given 100 studies that report results significant at $p \leq 0.05$, 5 should be irreproducible because they're false. This doesn't mean that the remaining 95 experiments will all be reproducible, however. That would be true only if the follow-up study were able detect all of the true results and it can't. Variability rears its head again. A statistical test has only a probability of identifying a true result when it occurs. Without going into details, the concept of *statistical power* rates the ability of a given experimental test design to detect genuine effects. Power varies from 0 to 1.0, where 0 means the test is wholly ineffective in finding true effects and 1.0 means the test is infallible; in practice a power level of 0.8 is considered to be very good.

Back to reproducibility. Less-than-perfect power implies that we must fail to reproduce a fraction of the true results and will incorrectly conclude that they are irreproducible. As applied to our remaining population of 95 true results (out of 100 and a p-value of 0.05), a follow-up study with a power of 0.8, could expect to reproduce only 76 (0.8 95) of them.

In sum, we anticipate being unable to reproduce 5 results because they are false and 19 (95-76) others because, although they're true, our tests are insufficiently powerful to reproduce them. Thus, statistical considerations alone suggest that, on average, we'll judge 24 of 100 studies to be irreproducible. And that number goes up as the power of our tests decreases. The anticipated irreproducibility of nearly one-quarter of published results has nothing to do with poor scientific practice, materials, or misconduct: it is baked into the math.

We reach a similar conclusion if, instead of p-values, significance levels, and statistical power, we use *effect sizes* and *confidence intervals*. The Reproducibility Project: Psychology (RPP) was an ambitious, meticulously planned, and diligently carried out effort to reproduce findings reported in 100 psychology papers. The RPP, often cited as hard evidence of a reproducibility crisis, employed several methods of assessing reproducibility and concluded that only approximately 40% of its attempts were successful. On the other hand, using a calculation based on effect sizes and confidence intervals, the RPP derived an expected reproducibility rate of 78.5%, implying that something closer to 50% of their attempts reproduced published findings.

Of course, the RPP was a scientific study, and as such, is itself the proper subject of scrutiny and criticism – metascience is still science! – and a debate about "crisis," "problem," or "nothing to worry about," continues. Nevertheless, it is generally agreed that there is a sizeable probability that a follow-up study will not succeed owing to factors inherent to science and statistics. Naturally, we can and should work hard to decrease the degree of irreproducibility that can be affected by improved practice. Still, a non-negligible, built-in failure rate has implications for understanding the reproducibility problem, as we'll see after looking into a few related factors.

Unforeseen impediments to reproducibility

Up to now, I've been assuming that optimal scientific procedures can be followed which include

controlling for relevant variables among conditions. In reality, it is impossible to control for all influential variables, partly because many of them are unknown. Two recent examples from the literature make this point. In one, observant workers in one animal research laboratory noticed that when female investigators did the experiment their results regularly differed from those obtained by male investigators who did the same experiment. Working through possible hypotheses to account for the differences (e.g., they found that having the individual investigators wear new t-shirts overnight and placing the shirts near the animals' cages caused the same divergence of results as the people themselves did), the lab eventually identified a male hormone that was triggering stress responses in the animals. If a group of female investigators had published their results, it is likely that a group of male investigators, or even a mixed male-female group, would not be able to reproduce them. (Must we now list the gender of the experimenter in the Methods sections of our papers?)

In a second example, two groups of investigators, one in Massachusetts and one in California, had a long-distance collaboration to study gene-expression in single cells from human breast cancer tissue. Frustratingly, the groups couldn't get the same baseline gene profiles although they were starting with the same tumour samples. After a year-long effort, they discovered that a seemingly trivial technical difference was responsible. Needing to free individual cells from tumour sections floating in saline solution, one group used a device that gently rocked the solution containing the sections back-and-forth, while the other used a device that swirled them in circles in a beaker. Nobody knows why the methods caused such different outcomes.

In both cases minor, easily overlooked variables had effects that could cause a lab's results to be irreproducible. Reacting to these stories, a colleague of mine remarked that such cases were "one in a million;" i.e. so rare that they could safely be ignored. Perhaps. However, we know about these two examples solely because the groups involved went out of their way to track down peculiar anomalies. In fact, nobody actually knows how common such occurrences are. Absent that information, the mere suspicion that they can happen will be enough to dilute scientists' enthusiasm for undertaking thorough reproducibility studies. Other factors may do the same.

Should I do a reproducibility study?

Reproducibility is a virtue of science, whereas irreproducibility is a problem for science and a major headache for individual scientists. Consider, what is the immediate, tangible value of a positive experimental replication? There are a few things we can say: When a follow-up study successfully reproduces a published result, it is undeniably reassuring and suggests that we may be on the right track. A true result must ultimately be reproducible. Of course, the converse – that a reproducible result must be true – is the fallacy of *affirming the consequent*. Karl Popper defines scientific *objectivity* as a form of intersubjective agreement; an observation that can, in principle, be made by anyone is an objective one. For him, unrepeatable observations cannot be tested by anyone, and therefore, are not objective. On the other hand, a reproduced result is not necessarily truer than it had been before it was reproduced; induction is as inconclusive as ever. Reproducibility helps to corroborate, though not to confirm, a hypothesis. A corroborated hypothesis can serve as the basis for practical action, even though further testing might show it

was wrong after all.

And there are many instances of empirical results that, although repeatedly reproduced, eventually turned out to be untrue. Newton's law of gravity is a textbook example of a theory that was tested and corroborated for hundreds of years, and yet, in the words of Nobel Prize winning physicist, Richard Feynman, is "just wrong." A wonderful and useful approximation at ordinary terrestrial scales; wrong at extremely large or small scales and at near-light speeds. Reproducibility alone is evidently not mandatory for progress. The real difficulty comes when a reproducibility trial fails; when a follow-up study does not reproduce the original. What conclusion does that lead to? Either the first or the second study could be wrong and the other right, both could be wrong or both right (with an unknown variable accounting for the differences). Apparently, the only things that are certain in this situation is that nature is more complicated than we had thought it was and that we need more data to sort things out. It is always good to be reminded of the vast intricacy of nature, but this is probably not the deep insight that is often assumed to follow from reproducibility studies. At first, an irreproducible result is just another problem to be solved. What's the next step? Try again? How many times is enough?

While reproducibility critiques tend to imply that every study should be reproduced, this cannot be done. Approximately 2.5 million science papers are published each year. Even if 90% were not high-quality original research reports, thoroughly reproducing approximately 0.25 million papers is hardly feasible either. Most papers relate the details of more than one experimental manipulation or test – in neuroscience, a reasonable estimate would be 3 – 10 experiments per paper – and the paper's overall conclusion is derived by aggregating the results of all of them. Recognising the enormous difficulties presented by multi-experiment papers, the authors of the aforementioned RPP selected just one result from each paper for their replication attempts. Despite seeming sensible, this strategy is problematic. How should investigators decide which component experiment to replicate? The RPP authors chose the last one while admitting that it might not be representative. More generally and significantly, does the irreproducibility of one experiment necessarily invalidate the conclusion of an entire report? It is also worth reiterating that the reproducibility decision ultimately rests on a conventional statistical benchmark for acceptance; any single result might be false at some level of probability. Yet, in the end, the truth (or falsity) of the unified conclusion of the whole study is more crucial for science than any single result. Thus far, reproducibility critiques do not seem to have taken up this issue.

While the preceding considerations do not entirely undermine the value of reproducibility, they do illustrate the many hazards in the way of achieving it. Mindful of this practical complexity, the individual laboratory supervisor faces a serious conundrum, namely, whether to attempt a replication study at all. Everyone who runs a lab must balance the uneven, and occasionally abstract, rewards of doing a reproducibility study against its all-too concrete up-front costs which, given typically limited funds, time, and personnel, may be significant. How much to invest when faced with a non-trivial chance that there might be a rational, though not very exciting, reason – e.g., species differences in biology – for any discrepancy that arises?

And, finally, let's not forget the calculations outlined earlier that suggest that 25-50% of otherwise

impeccably carried out experiments (at the $p < 0.05$ level) may fail for purely statistical reasons.

I suspect that these sorts of considerations, even if not explicit, help account for the reluctance of many scientists to take on thorough-going reproducibility studies. But then, say the reproducibility advocates, how can science progress? If findings are not reproduced, then what is the foundation of our trust in science?

If not reproducibility, what?

In spite of its value to science writ large, rigorously vetting someone else's results for reproducibility offers dubious benefits to those conducting science on a day-to-day basis. The reality of countless potential alternative explanations for an irreproducible result, constraints imposed by limited resources, and variable rewards attaching to reproducibility studies all militate against it. And reproducibility is not required for making progress. Science seeks Truth, for clear and complete explanations of nature. (In an Opinion piece in the August 2019 Issue of the HPS&ST NEWSLETTER, David Kennefick recounts the disappointing history of attempts to reproduce Eddington's measurement of the gravitational bending of light predicted by Einstein. The field moved on.)

Our confidence in a theory is often markedly increased by new observations that are consistent with it even prior to their replication. The recent detection of gravity waves and the photograph of a black hole, were welcomed as impressive corroborations of Einstein's theory even before they were reproduced.

Direct reproducibility *per se* is not the gold standard because science advances in the end by proposing, testing, rejecting, or refining, hypotheses. Stronger hypotheses force out weaker ones; the winnowing-out process can take a long time, and direct reproducibility plays a supporting role in it, but not the lead. A well-known example from neuroscience took place throughout the 1980s and was called, hyperbolically, "The LTP Wars." In this intellectual conflict, two opposing schools of thought clashed in their explanations of the cellular basis of learning in the mammalian brain. Cellular communication takes place via the diffusion of minute amounts of a chemical neurotransmitter across tiny junctions between neurons called synapses. Learning was believed to involve the physiological "strengthening" of synapses.. The LTP Wars were fought to determine whether the strengthening process took place at the signal-sending (pre-synaptic) side or the signal-receiving (post-synaptic) side of the synapse. The opposing hypotheses – pre- or post-synaptic – made entirely different predictions about what molecular changes formed the basis of learning. The Wars were contested by proposing and testing novel predictions of the two hypothesis. Staccato progress was made when one hypothesis accounted for observations that the other could not explain. After roughly a decade of contention, a kind of community consensus emerged that, at the synapses being studied, that the post-synaptic side won. Extensive, direct reproducibility studies never played a decisive role.

Thus, for many basic researchers, carrying out thorough reproducibility studies is normally on a back burner. When does reproducibility move to the front; when is it mandatory? There is no single set of circumstances, however, we should expect rigorously reproducible results in certain situations, especially those in which high costs – ethical, financial, societal – are involved. Examples include deciding to administer a therapeutic treatment to humans or incurring major

costs in pursuit of a technology affecting wide swaths of society. In other words, reproducibility is of greatest significance to applied science problems where definitive, all-or-none, actions must be taken; where the option to postpone a decision by seeking further, independent evidence does not exist, as it does in much of basic science.

Why do we trust in science?

Science needs to be assured of the explanatory soundness of a hypothesis: of its ability to account for an array of conditions related to the phenomena in question, of its generality to other similar conditions, of its ability to predict future phenomena, and of the precision and specificity of its explanations. In short, the kinds of information provided by experiments that have been misleadingly subsumed under the umbrellas of *systematic* and *conceptual* reproducibility, but which, as I've argued, are actually tests of hypotheses. This point is so often overlooked by the critics that it deserves emphasis: anxiety over a "reproducibility crisis," depends in part on a failure to recognise that testing and rejecting false hypotheses does not indicate a flaw in science. It is the very essence of the scientific method; it is a feature, not a bug. We trust in science to provide us the best understanding of the world given the limitations of our current knowledge because we are willing to jettison a worse explanation when a better one comes along. This is not to say that the decision to get rid of a worse, falsified, hypothesis is a simple one.

It is neither surprising nor irrational that scientists do not drop a well-corroborated hypothesis at the first sign of trouble. If we accept the dictum that no scientific fact can be established to be 100% true, we'd be foolish to do so. Science did not abandon Newton's theory of gravity because it could not obviously account for irregularities in the solar orbit of Uranus. Instead, the astronomer Urbain Le Verrier assumed Newton was right and correctly predicted that an unknown planet (which turned out to be Neptune) explained Uranus's orbit. As Popper cautioned, apparent falsification cannot be entirely certain either.

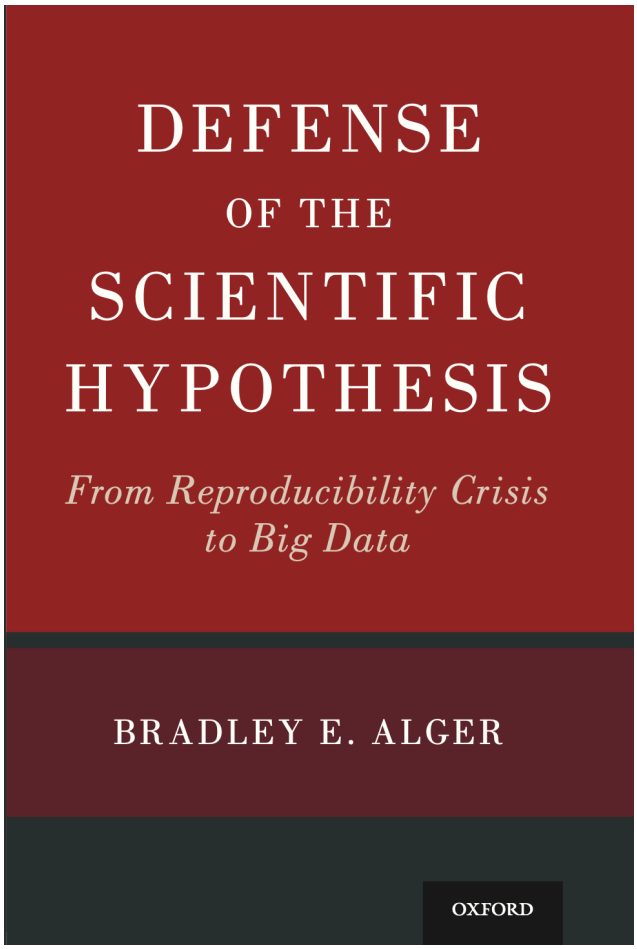
Be careful what you wish for: when insisting on reproducibility can be a problem. The concept that scientific confidence does not rest solely on reproducible results helps resolve another conundrum: how to evaluate investigations of events that cannot be reproduced? Geologists and geophysicists have done brilliant work in figuring out the phenomena that triggered the extinctions of the dinosaurs 66 or so million years ago. The leading contender, the hypothesis that an asteroid struck the earth near Chicxulub, Mexico, made predictions that have been tested and confirmed. Although dinosaur-extinction was a one-time-only event that precluded reproducibility testing in ideal detail, our present confidence in the hypothesis is high.

Indeed, an inflexible demand for reproducibility may backfire on those who support science. In 2015 Congress passed the Secret Science Reform Act that required federal agencies, including the Environmental Protection Agency (EPA), to base their decisions on the highest scientific standards; reproducibility, and its cousin "transparency," led the list of criteria of the best science. Despite its commendable emphasis on scientific evidence, some in Congress opposed bill because they feared it could be used to restrict the agencies' efforts. They were afraid that good science derived from fundamentally irreproducible studies would be excluded. For example, the catastrophic Deep Water Horizon oil spill in the Gulf of Mexico was

a singular event that yielded a trove of invaluable scientific information. Would such studies of environmental damage be off-limits? Would the EPA be forbidden to rely on decades-long studies of tobacco smokers unless the studies were fully reproduced? However unrealistic you might think these concerns are, they proved to be prophetic. In 2019, the EPA rolled back a number of critical environmental regulations, justifying its decisions by pointing, e.g., to the failure of long-term health studies to meet rigid reproducibility and transparency requirements (1).

Conclusions

Reproducibility is a multifaceted and intricate issue, an undisputed scientific virtue, but ultimately, not the most important one. Reproducibility is hard to define and, frequently, harder to achieve. In its search for true explanations for natural phenomena, science uses many standards and subgoals besides reproducibility; explanatory completeness – the ability of a hypothesis to account for a variety of observations, including non-obvious predictions, more effectively than other hypotheses; consistency with existing well-established theories; success in making predictions about future events, and quantitative precision, to name a few. While reproducibility strengthens conclusions driven by these intellectual concerns, it does not guarantee scientific validity, nor does its absence prevent progress. A better appreciation of reproducibility will foster a more realistic view of science and, in addition, will help us avoid mistakes that can come from overestimating its influence.



The above essay is primarily derived from Chapter 7 of:
Alger, Bradley, *Defense of the Scientific Hypothesis: From Reproducibility Crisis to Big Data*, Oxford University Press, Oxford. See also https//www.scientifichypothesis.org.

NOTE: It is anticipated that a subsequent Opinion Piece will address the question of how issues raised in this essay bear upon comparable concerns in education research.