Opinion Page

Creeping Bias in Research: Negative Results Are Glossed Over

New York Times, 24 September 2018

When we think of biases in research, the one that most often makes the news is a researcher's financial conflict of interest. But another bias, one possibly even more pernicious, is how research is published and used in supporting future work.

A recent study in *Psychological Medicine* examined how four of these types of biases came into play in research on antidepressants. The authors created a data set containing 105 studies of antidepressants that were registered with the Food and Drug Administration. Drug companies are required to register trials before they are done, so the researchers knew they had more complete information than what might appear in the medical literature.

Publication bias refers to the decision on whether to publish results based on the outcomes found. With the 105 studies on antidepressants, half were considered "positive" by the F.D.A., and half were considered "negative." Ninety-eight percent of the positive trials were published; only 48 percent of the negative ones were.

Outcome reporting bias refers to writing up only the results in a trial that appear positive, while failing to report those that appear negative. In 10 of the 25 negative studies, studies that were considered negative by the F.D.A. were reported as positive by the researchers, by switching a secondary outcome with a primary one, and reporting it as if it were the original intent of the researchers, or just by not reporting negative results.

Spin refers to using language, often in the abstract or summary of the study, to make negative results appear positive. Of the 15 remaining "negative" articles, 11 used spin to puff up the results. Some talked about statistically nonsignificant results as if they were positive, by referring only to the numerical outcomes. Others referred to trends in the data, even though they lacked significance. Only four articles reported negative results without spin.

Spin works. A randomized controlled trial found that clinicians who read abstracts in which nonsignificant results for cancer treatments were rewritten with spin were more likely to think the treatment was beneficial and more interested in reading the full-text article. It gets worse. Research becomes amplified by citation in future papers. The more it's discussed, the more it's disseminated both in future work and in practice. Positive studies were cited three times more than negative studies. This is **citation bias**.

Only half of the research was positive. Almost no one would know that. Even thorough reviews of the literature would find that nearly all studies were positive, and those that were negative were ignored. This is one reason you wind up with 10 percent of Americans on antidepressants when good research shows the efficacy of many of the drugs is far less than believed.

The preregistration of trials is supposed to help control for these biases. It works sporadically. In 2011, researchers examined cohorts of randomized controlled trials to see how well the published research matched what scientists said it was going to do beforehand. In some studies, they found, eligibility criteria for participants differed greatly from what was published.

In some, they found that procedures had changed for how to conduct analyses. In almost all, the sample size calculations had changed. Almost none reported on all the outcomes that were noted in the protocols or registries. Primary outcomes were changed or dropped in up to half of publications. This isn't to say secondary outcomes don't matter; they're often very important. It's also possible that some of these decisions were made for legitimate reasons, but, too often, there are no explanations.

In 2012, researchers re-analyzed 42 meta-analyses for nine drugs in six classes that had been approved by the F.D.A. In their re-analyses, they included data from the F.D.A. that was not in the medical literature. The addition of the new data changed the results in more than 90 percent of the studies. In those where efficacy went down, it did so by a median 11 percent. When efficacy went up – about the same rate that it went down – it did so by a median 13 percent.

This problem is worldwide. In 2004 in *JAMA*, a study reviewed more than 100 trials approved by a scientific-ethical committee in Denmark that resulted in 122 publications and more than 3,700 outcomes. But a great deal went unreported: about half of the outcomes on whether the drugs worked, and about two-thirds of the outcomes on whether the drugs caused harm. Positive outcomes were more likely to be reported. More than 60 percent of trials had at least one primary outcome changed or dropped.

But when the researchers surveyed the scientists who conducted the trials and published the results, 86 percent reported that there were no unpublished outcomes.

There has even been a systematic review of the many studies of these types of biases. It provides empirical evidence that the biases are widespread and cover many domains.

A modeling study published in *BMJ Open* in 2014 showed that if a publication bias caused positive findings to be published at four times the rate of negative ones for a particular treatment, 90 percent of large meta-analyses would later conclude that the treatment worked when it actually didn't.

This doesn't mean we should discount all results from medical trials. It means that we need, more than ever, to reproduce research to make sure it's robust. Dispassionate third parties who attempt to achieve the same results will fail to do so if the reported findings have been massaged in some way.

Further, there are things we can do to fix this problem. We can demand that trial results be published, regardless of findings. To that end, we can encourage journals to publish negative results as doggedly as positive ones. We can ensure that preregistered protocols and outcomes are the ones that are finally reported in the literature. We can hold authors to more rigorous standards when they publish, so that results are accurately and transparently reported. We can celebrate and elevate negative results, in both our arguments and reporting, as we do positive ones. Unfortunately, getting such research published is harder than it should be.

These actions might make for more boring news and more tempered enthusiasm. But they might also lead to more accurate science.

Comment: ***John Sweller***, School of Education, University of New South Wales

This is an important issue that is even more complex than the article suggests. I'm not clear what "negative results" mean in the context of instructional design.

Let me give examples from my Cognitive Load research history. Let's assume I run a simple, 2-group study hypothesising that Instructional Procedure A gives better test results than Instructional Procedure B. If my test results indicate that A is statistically better than B, the experiment will almost certainly be published. But if I get the opposite "negative" statistically significant result, it is equally likely to be published. Cognitive load theory is under constant development and those advances usually occur after such negative results.

The problem arises when we obtain a non-significant result. That is a negative result that is far less likely to be published. Our difficulty arises in determining why the non-significant result was obtained. The experiment may have failed because the hypothesis was wrong. That obviously should be published. But there are a multitude of trivial reasons why an instructional design experiment may fail and there is no point publishing an experiment if it fails for any of those reasons.

For example, the effect may be real but it may be too small to be detected by the statistical test. Repeating the experiment with a larger sample may yield significant results. There may be a mis-match between the materials used and the knowledge of the learners. If the learners are too advanced for the materials being taught, we'll get ceiling effects; if they are insufficiently expert we'll get floor effects. In either case, significant effects are unlikely and the experiment has obtained negative results for entirely trivial reasons.

There are many other trivial reasons for the failure of an experiment. I can't see the point of filling the literature with this stuff – there is enough useless stuff out there already. Notwithstanding these issues, I think the current system works. What doesn't work is non-research that can't possibly fail because of the way it is run.