Psychology's Woes and a Partial Cure: The Value of Replication

January 31, 2012

Henry L. Roediger, III

Psychology has come in for some bruising in the news media recently. The huge fraud case involving Diederik Stapel of Tilburg University in the Netherlands reaped a large amount of (well-deserved) negative publicity. Coming on the heels of other fraud cases at well-respected universities in North America, some see a trend emerging. In addition, recent publications provide additional ammunition for those firing at psychology in that the papers show that some researchers employ shoddy research practices (e.g., cherry-picking data to make some point) or use wrong statistics to bolster their claims (this kind of study comes along every few years, it seems).

The media had a number of different takes on the Stapel affair. The *Los Angeles Times* ran a story (by Amina Khan; November 5, 2011) headlined "Dutch scientist accused of falsifying data" that focused on Stapel and his fraud. Other media outlets were more generous with their blame. The *Chronicle of Higher Education* provided several stories. The first (on November 3) focused on Stapel ("The fraud who fooled almost everyone"). However, a later story on the front page of the *Chronicle* (November 13) ran under the headline "Fraud scandal fuels debate over practices of social psychology." Now it was not just Stapel who was under scrutiny, but the entire field of social psychology. The author, Christopher Shea, brought into play the article that was published in *Psychological Science* by Joseph Simmons, Uri Simonsohn and Leif Nelson on "false-positive psychology." These authors provided an interesting case study of how a complex piece of research can be done with many variables and then, if the authors cherry-pick their findings and ignore basic statistical practices, they can reach an outlandish conclusion (e.g., that listening to the Beatles "When I'm 64" can make people younger). The basic tactic was to use

all sorts of covariates in splitting data in various ways to find one that produced a significant outcome (father's age was used as a covariate in the case of the Beatles' song). The reach of the Stapel affair, the Simmons et al. article, and other matters (the cases of fraud involving Marc Hauser and, longer ago, Karen Ruggiero at Harvard) has worked others into a lather. Writing in the *New York Times* on November 2, Benedict Carey indicted all of scientific psychology in an article headlined "Fraud case seen as a red flag for psychology research."

Really? All of psychological research? Why not all of science? After all, the problems of fraud, nonreplication and poor statistical practices are hardly unique to psychology. As I was writing this piece, I read a notice of *Findings of Misconduct* from NIH, which reported several cases of scientific mischief (two involving plagiarism by computer scientists and another involving falsification of figures published in the *Journal of Cell Biology*). As this paper was going to press, another case of fraud involving a medical researcher at the University of Connecticut was in the news.

Calmer voices have been seeking to soothe these fevered proclamations about psychological science. APS Executive Director Alan Kraut wrote an op-ed piece arguing that "Despite Occasional Scandals, Science Can Police Itself" (in the *Chronicle of Higher Education*, December 9, 2011). Science is a self-correcting process, although sometimes the correction is slow in coming.

Social/personality psychologists have been agonizing over the events mentioned above, as well as publication of Daryl Bem's article on psi processes in the *Journal of Personality and Social Psychology*. Most psychologists I know have been paying close attention. I too have been reading through some of the commentaries about the whole business, trying to make sense of it all. The issues are varied and somewhat unrelated. Fraud is the most heinous of scientific crimes and can be hard to catch (although how Stapel carried on for so long is hard to fathom). Poor research practices as elaborated by Simmons et al. (2011) may be more common, but I think that someone who carried out a study similar to theirs (using many covariates and reporting only analyses that reached outlandish conclusions) would also border on fraudulent research.

What can psychology as a field do about these problems? Simmons et al. suggested six concrete steps that seem quite sensible (although "authors must collect 20 observations per cell" seemed a tad arbitrary). However, they omitted what, to me, seems the most obvious solution: replication of results. We should value replication more than we do, treasure it even. We were all routinely taught the value of replication in our first research methods course, but it seems some have forgotten the lesson. Of course, replication would not directly solve problems like the Stapel fraud — he could have made up several sets of data nearly as easily as he could have made up the first one. However, if others had tried to replicate his work soon after its publication, his misdeeds might have been uncovered much more quickly. Yet, my friends in social/personality psychology tell me that replication is often not encouraged or valued in their field. Writing an editorial in *Science*, Jennifer Crocker and Lynne Cooper (socal/personality psychologists) wrote "studies that replicate (or fail to replicate) others' findings are almost impossible to publish in top scientific journals." Brent Roberts (in a column in *P: The Online Newsletter for Personality Science*, which is published by the Association for Research in Personality) wrote: "In personality psychology, and most other areas of psychology, we actively devalue direct replication." He goes on to decry this tendency and to recommend steps to remedy it.

I am not sure replication is always devalued. At least in my little corner of the world of psychological

science, I see replications all the time. Often, for cognitive psychologists, replications of experiments are required for publication by editors in our most prestigious journals. While in graduate school, I was admonished repeatedly on the critical importance of replication and was taught to never ever submit a finding that you were not sure of via replication. To those who argue that a robust level of statistical significance is all one needs to assure replicability, I recall the aphorism (attributed to Confucius) that "One replication is worth a thousand *t*-tests." Words to live by. And if we replicate our results routinely, we do not need to worry so much about the poor logic of null hypothesis statistics or using Bayesian statistics to try to determine what happened in a single experiment or study. If you obtain an effect, just replicate it (perhaps under somewhat different conditions) to be sure it is real. I will illustrate the benefits of replication with a personal example below.

A Tale of Two Studies

How can we avoid the problem of nonreplication that seems to plague psychological science and other fields? The answer is disarmingly simple: Researchers should always, whenever possible, replicate a pattern of results before publishing it. The phenomenon of interest should be subjected to careful scrutiny, should be twisted, bent, and hammered to see if it will survive. If the basic effect is replicated under the exact conditions as in the original study, but it disappears when conditions are changed a bit, then the effect is real but brittle; the boundary conditions for obtaining the effect are rather narrow. That is not ideal, but is certainly worth knowing. Many phenomena in the world of cognitive psychology have this feature of holding under one set of conditions (say, in within-subject designs) but disappearing under another set of conditions (in between-subject designs). McDaniel and Bugg (*Psychonomic Bulletin & Review*, 2008) review how many interesting memory phenomena (even strong ones, like the generation effect) can be affected by the type of design employed. That is simply a fact that would need to be explained, but not a failure to replicate, at least in one sense (to be discussed further below).

In the mid-1990s, Kathleen McDermott and I were collaborating on research, and we tried two rather risky experiments, ones that seemed likely to fail but that were worth trying. To our surprise, we found startling patterns of data in both procedures. Yes, in both cases, we found what we predicted (or at least what we hoped to find), but we were skeptical about the results.

One case involved a technique for studying false memories in a list-learning situation in which the illusory memories seemed to occur nearly immediately and to be remarkably strong (contrary to standard paradigms of the time used to induce false memories). After a first classroom pilot experiment, we conducted a proper second experiment that confirmed and strengthened our initial results. We started to write up the two experiments. However, we were still a bit worried about the robustness of the effects, so we continued experimenting while we wrote. We were able to confirm the results in new experiments (employing various twists), so that by the time the paper reporting two experiments was accepted and published in the *Journal of Experimental Psychology: Learning, Memory and Cognition* in 1995, we had several more replications and extensions ready to be written.

Our paper had been fairly widely circulated as a preprint and generated some excitement in our little research world, so soon after publication I began getting manuscripts to review that used the same technique. The papers all began with a basic replication of our effect (although sometimes the replication was presented as a control condition to be contrasted with other conditions). Why? I suspect the answer was that the other researchers disbelieved our results or were at least skeptical, so they wanted to

demonstrate the effect for themselves before exploring it. These papers replicating and extending the associative-list false memory effect were quickly published — no problem in getting replications published in this instance — and thus, within two years of its initial publication, anyone in my field who cared could know that the effect reported by Roediger and McDermott (1995) was genuine. (The basic effect has now been replicated hundreds of times.) Yes, McDermott and I had replicated our basic effect in the original paper, but the fact that others confirmed it many times over was critical to establishing it as genuine.

The second experiment we were excited about at that time did not have so happy a fate. Briefly, we developed a new (or newish) technique for measuring recognition memory that we suspected (from literature in animals) might be more sensitive than the usual tests of recognition memory. (I will skip the details for reasons that will become obvious, if they are not already.) Our first experiment manipulated a standard variable and measured recognition using our new method and a standard method. To our delight, we found that the new recognition method was indeed more sensitive than the old one (there was a main effect favoring it and an interaction as a function of another variable, showing that the new method was more sensitive than the standard one). We were elated; the effects were quite reasonable, the statistics were robust, and we were off to the races. Or so we thought. We felt confident enough to submit the research to be presented as a talk at the Psychonomic Society annual meeting, and the work was presented in 1995 in Los Angeles.

After the talk, we decided we needed to replicate and extend the effect, to make sure it was replicable and robust, before submitting it for publication. So we tried replicating the experiment with a twist (a new independent variable), a new subject population (undergraduates, because the original experiment had been done with Air Force recruits), but with the same two measures of recognition memory (standard and new). We got a pattern that looked slightly hopeful, but was far from being statistically significant; we deemed it a failure to replicate (or at least certainly not a success). We scratched our heads and tried again. For the third experiment, we went back to the exact design and procedure to try a direct replication of the method and procedure, albeit still with undergraduates. Again, we did not get the effect, and now the data looked terrible — no hint of an effect of the test variable (the standard versus new procedure) was obtained. (It might not be possible to prove the null hypothesis, but it certainly can be hard to reject it.) As noted, our original experiment had been with Air Force recruits and the next two were with undergraduates. Although we could not imagine a reason that the subject population should matter, we decided to try a direct replication at the Air Force base using more subjects than we had in the original experiment. We still could not get the effect; just null results. The two procedures seemed equivalent measures of recognition. Altogether, we tried several more times over the next few years to replicate the effect. To make a long story short, we never got it again, even though our original experiment in the series had produced such pretty results. Sometimes we got results that hinted at the effect in our new experiments, but more often the results glared out at us, dull and lifeless, telling us our pet idea was just wrong. We gave up.

McDermott and I might well have published our initial single initial experiment as a short report. After all, it was well conducted, the result was novel, we could tell a good story, and the initial statistics were convincing. I would bet strongly we could have had the paper accepted. Luckily, we did not pollute the literature with our unreplicable data — but only because we required replication ourselves (even if the editors probably would not have — brief reports do not encourage and sometimes do not permit replication).

The moral of the story is obvious: Replicate your own work prior to publication. Don't let others find out that you are wrong or that your work is tightly constrained by boundary conditions. If there were a way to retract conference papers, we would have retracted that one. Most people don't count conference presentations as "real" for the scientific literature, and our case provides another good reason for that attitude. At least we found out that our effect was not replicable before we published it.

Varieties of Replication

Nearly every research methods textbook harps on the need for replication. In my experience, it is fairly easy to get successful replications of work published because usually the replication is presented in the context of other research that extends the basic phenomenon of interest. On the other hand, failures to replicate are much more difficult to publish. This fact is bemoaned, but in a way is as it should be. For someone to claim a "failure to replicate" someone else's work, the person needs to have really tried hard to do so. A one-shot "we-tried-but-didn't-get-it" attempt is not enough. Some failures to replicate are published, at least within cognitive psychology (see Fernandez & Glenberg, *Memory & Cognition*, 1985, for one paradigmatic case study in how to conduct and publish a failure to replicate).

The concept of replication is often treated as well defined and unitary. You replicate or you do not. Of course, that is not so; it is customary to distinguish among several types of replication attempts: direct replication, systematic replication, and conceptual replication. As the name implies, direct replications attempt to reproduce a result using the same conditions, materials and procedures as in the original publication to make a replication as close as possible to the original research. Systematic replications are an attempt to obtain the same finding, but under somewhat different conditions (say, in a memory experiment, with a different set of materials or a different type of test). Finally, a conceptual replication tries to replicate the existence of a concept (e.g., cognitive dissonance) by using a different paradigm (say, moving from an induced compliance paradigm for studying dissonance to a free choice paradigm). If the researcher cannot find evidence of cognitive dissonance in the latter paradigm, the result has no necessary implication for replicability of the experiment using the former paradigm. Of course, both these paradigms have been frequently shown to produce cognitive dissonance in line with the core idea of the concept.

When someone uses the phrase "failure to replicate," they almost always have in mind (or should have in mind) direct replication. However, even the concept of direct replication represents a continuum. For example, it is never possible to test the same subjects from the original study, nor is it possible to use the same equipment. Thus, one must make judicious judgments about how close is close enough, and in my experience, debates between the authors of an original report and those trying to publish a failure to replicate it often differ on what "close enough" means. Often, a replication attempt will use the same number of subjects as in the original attempt. This approach sounds reasonable, but studies have shown that experiments will often fail to replicate using this strategy (even if the effect is real). So it would be wise to use 150 percent or more of the number of subjects in the original, if possible.

Is the solution to scientific psychology's woes as simple as replication? Well, no, or at least not completely. However, I would argue that by following the practice of both direct and systematic replication, of our own research and of others' work, we would avoid the greatest problems we are now witnessing. In truth, this advice is easier to adopt for some fields than for others. In most types of cognitive research, replication is fairly easy. But in some types of research (those with special

populations, or onerous manipulations, or longitudinal studies) are by definition difficult to replicate. In these cases, we must depend on other scientific tactics to insure validity of the study. Nonetheless, much of scientific psychology is composed of the sorts of studies that can be readily replicated with just a bit more work, and the replications can be of the systematic variety (changing things up a bit) rather than simply direct replications. We must also ask editors to be open to devoting journal space to replications. In fact, reviewers and editors might be strongly encouraged to ask authors about replicability of their work (even if they are submitting a brief report with only a single study). Do they have other data for possible future reports that insure that the effects reported are genuine? The need for such assurance is particularly high when a single study reports some dramatic or surprising claim.

In a preceding paragraph, I wrote that it should be difficult to publish failures to replicate, in part because failures to replicate can be due to sloppiness on the part of the replicator rather than the original researchers. I implied that the onus should fall on the replicator to directly replicate the research, trying hard to do so in a systematic series of studies. Hal Pashler, who read an earlier version of this column, said he agreed with the general point, but he commented: “However, from a systemic point of view, it [the practice of not publishing failures to replicate unless they are exceptionally systematic] guarantees a biased scientific literature, because let’s face it, most of the time when people fail to replicate a result, one or two studies is all they bother to do.”

He went on to say that often these are students looking for a topic on which to do a thesis or dissertation, and faced with a failure to replicate, they are likely to give up and move onto another topic rather than to pursue a failure-to-replicate dissertation. Such a dissertation might be good for the field, but might not help the student on the job market. To quote Pashler’s note a bit further: “So if the typical program of research that yields a nonreplication isn’t ever taken to the point where we would say it should be considered worthy of publication, then errors in the literature will only rarely be corrected, and our literature (even our textbooks) will become bigger and bigger heaps of unreplicable junk (exactly as John Ioannidis’ famous 2005 paper in *PloS Medicine* led us to expect).”

Assuming this analysis is correct (and there is a sad ring of truth to it, if we do not replicate our work and that of others), what is the solution? Happily, Pashler is on the front line of providing a possible solution. Along with Barbara Spellman, Alex Holcombe, and Sean Kang, Pashler has helped to create a website called PsychFileDrawer.org to remedy the situation. As the website says, PsychFileDrawer.org is intended to be “An archive of replication attempts in experimental psychology.” The use of “experimental psychology” in this context is meant to be broad, to encompass all of scientific psychology. The authors of the website write that “PsychFileDrawer.org is a tool designed to address the file drawer problem as it pertains to psychological research: the distortion in the scientific literature that results from the failure to publish non-replications.” Its creators urge researchers to post their replication attempts (successful or unsuccessful) on the website, and the site specifies the rules of the game. I urge readers to take a look. Although the website has only recently gotten started, it should prove to be a useful addition for all fields of psychology. If several researchers (or research groups) report that they cannot replicate a particular finding, this would serve notice to the field that a “false-positive” result may exist in the literature. The various authors of the failures may even team up to publish their findings in an archival journal.

The recent critical examination of our field, though painful, may lead us to come out stronger on the other side. Of course, as noted above, failures to replicate and the other problems (fraud, the rush to

publish) are not unique to psychology. Far from it. A recent issue of *Science* (December 2, 2011; Volume 334, No. 6060) contained a section on “Data replication & reproducibility” that covered issues in many different fields. In addition, an article in the *Wall Street Journal* (“Scientists’ Elusive Goal: Reproducing Study Results,” December 2, 2011) covered failures to replicate in medical research. So, failures to replicate are not only a problem in psychology. Somehow, though, when an issue of fraud or a failure-to-replicate occurs in (say) field biology, journalists do not create headlines attacking field biology or even all of biology. It seems that psychology is special that way.