CREDITS (TOP TO BOTTOM) (COLLAGE): ILYA ANDRIYANOV/SHUTTERSTOCK IMAGES AND JOURNAL COVER APPEARS WITH PERMISSIONS FROM ASSOCIATION FOR PSYCHOLOGICAL SCIENCE; UNIVERSITY OF VIRGINIA

Psychology's Bold Initiative

In an unusual attempt at scientific self-examination, psychology researchers are scrutinizing their field's reproducibility

PICK UP THE JANUARY 2008 ISSUE OF *Psychological Science*, turn to page 49, and you'll find a study showing that people are more likely to cheat on a simple laboratory task if they have just read an essay arguing that free will is an illusion. It was a striking study that drew widespread attention both from psychologists and from many media outlets. But should you believe the result?

There's no reason to think that the study, conducted by psychologists Kathleen Vohs of the University of Minnesota Carlson School of Management in Minneapolis, and Jonathan Schooler, who is now at the University of California, Santa Barbara (UCSB), is incorrect. Yet according to many psychologists, their field has a credibility problem at the moment, and it affects thousands of studies like this one.

Part of the angst stems from recent high-profile cases of scientific misconduct, most dramatically the extensive fraud perpetrated by Dutch social psychologist Diederik Stapel (*Science*, 4 November 2011, p. 579), that have cast a harsh light on psychological science. Yet there is no evidence that psychology is more prone to fraud than any other field of science. The greater concern arises from several recent studies that have broadly critiqued psychological research practices, highlighting lax data collection, analysis, and reporting, and decrying a scientific culture that too heavily favors new and counterintuitive ideas over the confirmation of existing results. Some psychology researchers argue that this has led to too many findings that are striking for their novelty and published in respected journals—but are nonetheless false.

As a step toward testing that disturbing idea, one project begun this year offers an online site (PsychFileDrawer.org) where psychologists can quickly and easily post, in brief form, the results of replications of experiments—whether they succeed or fail. University of California, San Diego, psychologist Hal Pashler, one of the project's developers, says the goal is to counteract the "file drawer problem" that plagues all of science, including psychology; researchers usually just file away straightforward replication studies because most journals decline to publish such work.

Double trouble? Brian Nosek leads a large-scale effort to replicate recent psychology studies

In an even more daring effort, a group of more than 50 academic psychologists, which calls itself the Open Science Collaboration (OSC), has begun an unprecedented, large-scale project to systematically replicate psychological experiments recently published in leading journals. "We're wringing our hands worrying about whether reproducibility is a problem or not," says psychologist Brian Nosek of the University of Virginia in Charlottesville, who is coordinating the effort. "If there is a problem, we're going to find out, and then we'll figure out how to fix it."

Robert Kail, a Purdue University developmental psychologist and editor of *Psychological Science*—one of the three journals whose papers the OSC is attempting to replicate—is optimistic that a high percentage of published findings will be replicated. Nonetheless, he views the field's recent attention to the issue of false positives as healthy. "There has been a lot of speculation about the extent to which it's a problem," he says. "But nobody has actually set it up as an empirical project. It's a great thing for somebody to actually do that."

Schooler, who is not directly involved with the project but whose free will study

Published by AAAS

will be replicated, considers the OSC replication study a "bold initiative." Yet he's concerned that if the project confirms few studies, it could unfairly indict psychology. "I think one would want to see a similar effort done in another area before one concluded that low replication rates are unique to psychology," he says. "It would really be a shame if a field that was engaging in a careful attempt at evaluating itself were somehow punished for that. It would discourage other fields from doing the same."

Indeed, the prospect of exposing psychology's foibles has upset some scientists. "I had a senior person in the field ask me not to do it, because psychology is under threat and this could make us look bad," Nosek says. "I was stunned." The point of the project, he says, is not to single out individual studies or disparage psychology. "We're doing this because we love science. The goal is to align the values that science embodies—transparency, sharing, self-critique, reproducibility—with its practices."

Am I doing something wrong?

Reproducibility is supposedly a basic tenet of science, but a number of fields have raised concerns that modern publishing pressures inhibit replication of experiments. In a well-known 2005 *PLoS Medicine* essay, epidemiologist John Ioannidis, now at Stanford University in Palo Alto, California, argued that in biomedicine, many if not most published research claims are false. He outlined a number of factors—including small sample sizes, small effect sizes, and "flexibility" in the research process—that contribute to a high rate of false positives. One reason those false positives aren't caught is because of a lack of emphasis on replication studies, which is "standard across fields of science," says Columbia University statistician Victoria Stodden, who studies reproducibility and openness in computational science.

Nosek first became interested in the problem of replication in psychology several years ago, after he started having persistent problems confirming others' results—something his lab routinely does before extending research to address new questions. Believing himself to be a careful methodologist, he wondered whether something was wrong with the studies he was attempting to reproduce.

For example, many social psychological studies have asked participants to unscramble sentences containing words that connote a particular concept that the researchers aim to "prime" in the participants' minds—say, impulsivity, or anger, or happiness; the idea is to test the effects of those mental constructs on a subsequent task. The scrambled-sentence task has been used in many published studies, but Nosek's group has rarely been able to get it to work, and he suspects that the effect may be much more limited than the published literature would suggest.

The idea of systematically testing the reproducibility of psychological science percolated in Nosek's mind until late last year, when revelations of Stapel's scientific misconduct brought the issue to a boil. Stapel, whose studies were widely cited and had drawn frequent media attention, admitted last fall to fabricating data on more than 30 studies dating back to the mid-1990s. One of those high-profile studies, published in *Science* and now retracted, concluded that chaotic physical surroundings promote stereotyping and discrimination.

News of Stapel's fraud led many psychologists to question whether the field possesses sufficient checks and balances to prevent such willful deception. It also stirred some researchers' nascent worries that—rare acts of fraud aside—commonly accepted practices in psychology research might produce an unacceptably high number of false positive results. "It just became obvious that it was time to do some science to figure it out," Nosek says.

In November 2011, Nosek approached a few departmental colleagues to propose a collaborative effort to study the field's reproducibility. The effort soon expanded to include researchers from universities in the United States, Canada, and Europe.

This winter, OSC members began identifying studies to include in their replication sample. They chose three high-impact psychology journals—*Psychological Science*, the *Journal of Personality and Social Psychology*, and the *Journal of Experimental Psychology: Learning, Memory, and Cognition*—and began logging key details of the first 30 articles published in each journal in 2008. They reasoned that articles published during this time frame are recent enough that most original authors can find and share their materials, yet old enough for the OSC to analyze questions such as whether a study's reproducibility correlates with how often it has been cited subsequently. Many psychology articles include more than one study to support their conclusions, but the OSC decided in advance to select only the final study from each article; if that study was not feasible to replicate, they worked backward until an eligible study was identified.

After identifying eligible experiments, individual OSC members began choosing the ones each would repeat and contacting the original authors to obtain materials and fill in methodological details. So far, the response from original authors has been overwhelmingly positive—only two have declined to provide materials needed for replication.

More than 30 replication studies are now under way, and the group aims to conduct at least 50. (New researchers can join the collaboration anytime.) The results OSC members will seek to reproduce vary widely, from a study of how photos of unsmiling black men automatically activate attention in some people's minds, to a study that examined how the timing in which perceptual stimuli are presented affects short-term memory.

The replication studies are being funded by the OSC researchers who select them; however, in keeping with the samples used in the original studies, most will be cheap because they'll use undergraduates who participate for course credit—although using study populations that are mostly limited to young, white college students from Western industrialized cultures has its own drawbacks (*Science*, 25 June 2010, p. 1627).

Repeat after me

While reproducibility is often held up as the "gold standard" in science, Nosek argues that "direct replications," in which researchers follow an original experiment's procedure as closely as possible, are rare. "There is no incentive for replication—it's all about the new idea," he says.

Not all psychologists agree. Social psychologist Norbert Schwarz of the University of Michigan, Ann Arbor, says that although the OSC project is valuable, concern about the field's robustness may be overblown. Direct replications may be rare, but Schwarz points out that conceptual replications—in

Online
sciencemag.org
Podcast interview (http://scim.ag/pod_6076) with author Siri Carpenter.

> *"IT WILL BE A* ***FIRST ESTIMATE*** *OF THE* ***REPRODUCIBILITY OF FINDINGS*** *THAT ARE IN IMPORTANT JOURNALS IN PSYCHOLOGY."*
> **—BRIAN NOSEK, UNIVERSITY OF VIRGINIA**

Published by AAAS

which researchers tweak a study's materials or procedure to test a hypothesis in a different way—make up the bulk of psychology's literature and "are highly valued because they look at the stability of a phenomenon across different content domains."

Nosek agrees that conceptual replications are important and common, and temper worries that journals are littered with false findings. But conceptual replication assumes that the replication addresses the same phenomenon as the original demonstration, which may not always be the case, he says.

Direct replications are important, he and others say, because psychology, like other disciplines, has practices that may lead to too many false positives. In the November 2011 issue of *Psychological Science*, University of Pennsylvania psychologist Joseph Simmons and colleagues showed, through computer simulations and actual experiments, that "flexibility" in research decisions such as how many research subjects to include in a study, how many outcomes to measure, or whether to break down analyses according to participants' gender can more than double the chances of getting a false positive. When several such "researcher degrees of freedom" are in play, as is commonly the case, a study is more likely than not to mistakenly yield a statistically significant effect.

Nosek believes the prevalence of research practices that lead to unintended bias is rooted in the fact that negative results are virtually unpublishable—a state of affairs that has grown more extreme in recent years, especially in psychology. "It is important to my career that I do studies that are publishable," he says. One way to do that is to capitalize on the kinds of practices that Simmons and colleagues identified. Another is to run many studies with small numbers of subjects and publish the ones that "work." A third and especially insidious problem, Nosek says, is that researchers can easily fool themselves into believing that chance positive results are actually what they had hypothesized all along—then publish such findings as though they were confirmatory tests of existing hypotheses. There is currently nothing to prevent such "motivated reasoning," as psychologists call it, from contaminating the scientific literature.

Looking ahead

Nosek isn't just depending on OSC's look at past studies to improve psychology research practices. He's also looking forward and has developed, with his graduate student Jeffrey Spies, an online database (openscienceframework.org) where psychological scientists can easily organize and—if they choose—register study materials, hypotheses, planned analyses, and data. Nosek hopes that in addition to aiding laboratory workflow and providing an outlet for results that might not otherwise see the light of day, the registry will increase researchers' sense of accountability both to their community and to themselves. Looking forward to using the registry for his own research, Nosek says he sees it as "a way of guarding the truth against my own motivated reasoning."

> *"IT WOULD REALLY* ***BE A SHAME*** *IF A FIELD THAT WAS ENGAGING IN A CAREFUL ATTEMPT AT EVALUATING ITSELF WERE* ***SOMEHOW PUNISHED*** *FOR THAT. IT WOULD* ***DISCOURAGE OTHER FIELDS*** *FROM DOING THE SAME."*
>
> **—JONATHAN SCHOOLER, UNIVERSITY OF CALIFORNIA, SANTA BARBARA**

As for the OSC's replication project, data collection on the currently chosen studies should be completed by the end of this year, and Nosek hopes the results will be published not long after. One challenge for the OSC researchers will be in setting criteria for what constitutes a successful confirmation, given that a replication can take varying forms, from a result that has the same statistical significance as the original finding to a pattern of results that is merely in the same direction as the original. That makes it difficult to say what percentage of failed replications should be considered problematic. And it's one reason that Nosek believes the most interesting results of the study may lie not in the raw rate of reproducibility but in what factors predict a study's reproducibility.

For example, if the studies that replicate best are also the most widely cited, that would suggest that despite biases in publishing practices, scientists can separate the wheat from the chaff. If studies that were published along with several others in a multistudy paper—a formulation that often involves multiple confirmations of the same basic effect, at least conceptually—replicate more often than those that were single-study papers, that might hint that "bite size" reports are problematic, as some critics have suggested.

One limitation in interpreting the OSC's results stems from the fact that the group is not tackling a representative sample of all psychology research. Studies that used statistical analyses that cannot yield clear evidence of replication have been excluded, as have studies that would be infeasible to replicate because they require specialized materials, instrumentation, or participant populations. Furthermore, most studies included in the sample are drawn from cognitive and social psychology; other subfields, such as developmental and clinical psychology, neuroscience, and animal behavior, are not included.

Stanford University social psychologist Nalini Ambady says several junior colleagues have told her they're worried about this disproportionate focus because if a high percentage of OSC studies fail to be replicated, many people may conclude that it is social psychology alone that is problematic. She sympathizes with that argument. "I think if you want to do it, then you should do a fair, representative sampling," Ambady says. "Don't just focus on the social psychological studies that are the low-hanging fruit because they are generally cheapest and easiest to conduct."

The study's heavy focus on social and cognitive psychology is a reflection of the researchers who have gotten involved so far, Nosek responds: "If there are people in other subfields who want to join the project, we will be delighted to broaden it." The OSC won't produce a "definitive study," he stresses. "It will be a first estimate of the reproducibility of findings that are in important journals in psychology."

Columbia's Stodden agrees that psychology's efforts to address the issue shouldn't be cause for criticism. Psychologists' scrutiny "is very admirable," she says. "I think other fields could benefit from that kind of self-reflection."

Cognitive psychologist Rebecca Saxe of the Massachusetts Institute of Technology in Cambridge, who is participating in the collaboration, is also optimistic. The project, she says, "has the potential to be spun as negative and nihilist. But to me, it's the opposite. Science is about discovering true things, and when you do find something that's true enough that others are able to replicate it, that's just thrilling."

–SIRI CARPENTER

Siri Carpenter, a freelance writer based in Madison, Wisconsin, worked 12 years ago in a lab with Brian Nosek, the organizer of the Open Science Collaboration.