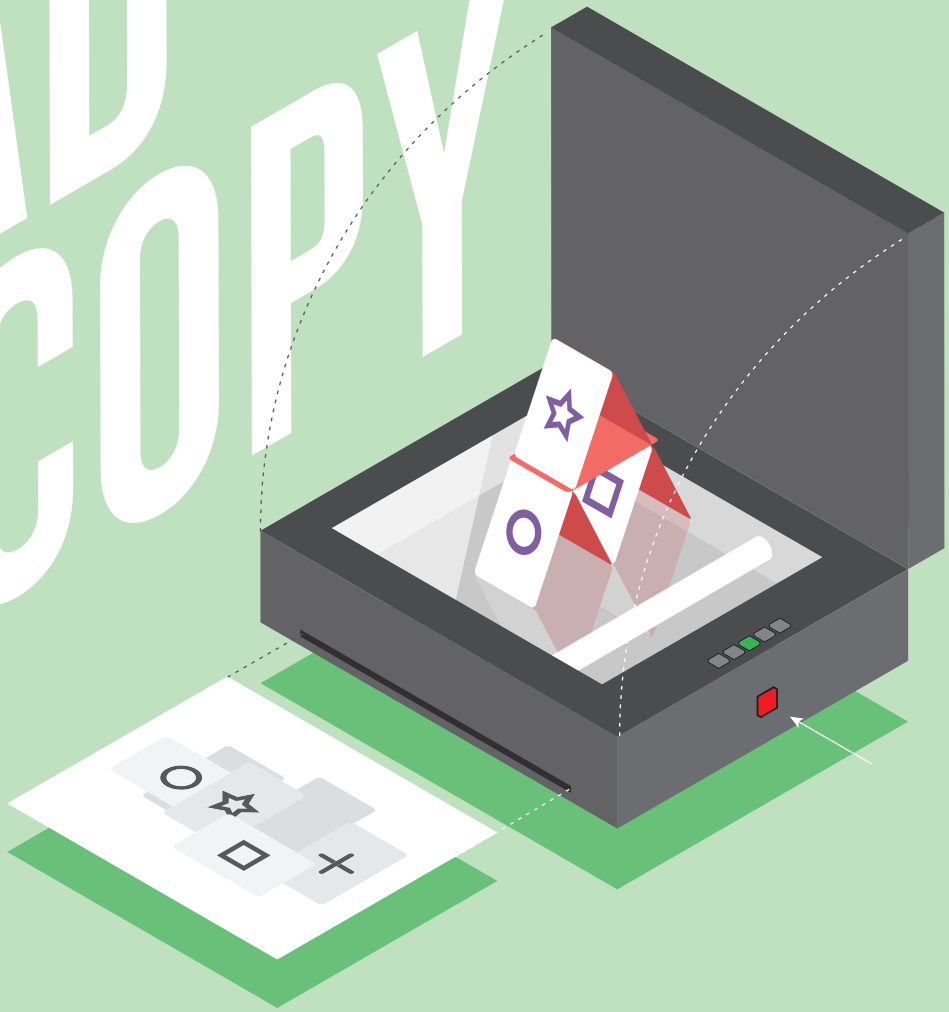
BAD COPY



IN THE WAKE OF HIGH-PROFILE CONTROVERSIES, PSYCHOLOGISTS ARE FACING UP TO PROBLEMS WITH REPLICATION.

BY ED YONG

JASIEK KRZYSZTOFIAK

For many psychologists, the clearest sign that their field was in trouble came, ironically, from a study about premonition. Daryl Bem, a social psychologist at Cornell University in Ithaca, New York, showed student volunteers 48 words and then abruptly asked them to write down as many as they could remember. Next came a practice session: students were given a random subset of the test words and were asked to type them out. Bem found that some students were more likely to remember words in the test if they had later practised them. Effect preceded cause.

Bem published his findings in the *Journal of Personality and Social Psychology* (*JPSP*) along with eight other experiments[1] providing evidence for what he refers to as "psi", or psychic effects. There is, needless to say, no shortage of scientists sceptical about his claims. Three research teams independently tried to replicate the effect Bem had reported and, when they could not, they faced serious obstacles to publishing their results. The episode served as a wake-up call. "The realization that some proportion of the findings in the literature simply might not replicate was brought home by the fact that there are more and more of these counterintuitive findings in the literature," says Eric-Jan Wagenmakers, a mathematical psychologist from the University of Amsterdam.

Positive results in psychology can behave like rumours: easy to release but hard to dispel. They dominate most journals, which strive to present new, exciting research. Meanwhile, attempts to replicate those studies, especially when the findings are negative, go unpublished, languishing in personal file drawers or circulating in conversations around the water cooler. "There are some

experiments that everyone knows don't replicate, but this knowledge doesn't get into the literature," says Wagenmakers. The publication barrier can be chilling, he adds. "I've seen students spending their entire PhD period trying to replicate a phenomenon, failing, and quitting academia because they had nothing to show for their time."

These problems occur throughout the sciences, but psychology has a number of deeply entrenched cultural norms that exacerbate them. It has become common practice, for example, to tweak experimental designs in ways that practically guarantee positive results. And once positive results are published, few researchers replicate the experiment exactly, instead carrying out 'conceptual replications' that test similar hypotheses using different methods. This practice, say critics, builds a house of cards on potentially shaky foundations.

These problems have been brought into sharp focus by some high-profile fraud cases, which many believe were able to flourish undetected because of the challenges of replication. Now psychologists are trying to fix their field. Initiatives are afoot to assess the scale of the problem and to give replication attempts a chance to be aired. "In the past six months, there are many more people talking and caring about this," says Joseph Simmons, an experimental psychologist at the University of Pennsylvania in Philadelphia. "I'm hoping it's reaching a tipping point."

PERVASIVE BIAS

Psychology is not alone in facing these problems. In a now-famous paper[2], John Ioannidis, an epidemiologist currently at Stanford School of Medicine in California argued that "most published research findings are false", according to statistical logic. In a survey of 4,600 studies from across the sciences, Daniele Fanelli, a social scientist at the University of Edinburgh, UK, found that the proportion of positive results rose by more than 22% between 1990 and 2007 (ref. 3). Psychology and psychiatry, according to other work by Fanelli[4], are the worst offenders: they are five times more likely to report a positive result than are the space sciences, which are at the other end of the spectrum (see 'Accentuate the positive'). The situation is not improving. In 1959, statistician Theodore Sterling found that 97% of the studies in four major psychology journals had reported statistically significant positive results[5]. When he repeated the analysis in 1995, nothing had changed[6].

> "TO SHOW THAT 'A' IS TRUE, YOU DON'T DO 'B'. YOU DO 'A' AGAIN."

One reason for the excess in positive results for psychology is an emphasis on "slightly freak-show-ish" results, says Chris Chambers, an experimental psychologist at Cardiff University, UK. "High-impact journals often regard psychology as a sort of parlour-trick area," he says. Results need to be exciting, eye-catching, even implausible. Simmons says that the blame lies partly in the review process. "When we review papers, we're often making authors prove that their findings are novel or interesting," he says. "We're not often making them prove that their findings are true."

Simmons should know. He recently published a tongue-in-cheek paper in *Psychological Science* 'showing' that listening to the song *When I'm Sixty-four* by the Beatles can actually reduce a listener's age by 1.5 years[7]. Simmons designed the experiments to show how "unacceptably easy" it can be to find statistically significant results to support a hypothesis. Many psychologists make on-the-fly decisions about key aspects of their studies, including how many volunteers to recruit, which variables to measure and how to analyse the results. These choices could be innocently made, but they give researchers the freedom to torture experiments and data until they produce positive results.

In a survey of more than 2,000 psychologists, Leslie John, a consumer psychologist from Harvard Business School in Boston, Massachusetts, showed that more than 50% had waited to decide whether to collect more data until they had checked the significance of their results, thereby allowing them to hold out until positive results materialize. More than 40% had selectively reported studies that "worked"[8]. On average, most respondents felt that these practices were defensible. "Many people continue to use these approaches because that is how they were taught," says Brent Roberts, a psychologist at the University of Illinois at Urbana–Champaign.

All this puts the burden of proof on those who try to replicate studies — but they face a tough slog. Consider the aftermath of Bem's notorious paper. When the three groups who failed to reproduce the word-recall results combined and submitted their results for publication, the *JPSP*, *Science* and *Psychological Science* all said that they do not publish straight replications. The *British Journal of Psychology* sent the paper out for peer review, but rejected it. Bem was one of the peer reviewers on the paper. The beleaguered paper eventually found a home at *PLoS ONE*[9], a journal that publishes all "technically sound" papers, regardless of novelty.

"I've done everything possible to encourage replications," says Bem, who stands by his results, and has put details of all his methods and tests online. But he adds that one replication paper is uninformative on its own. "It's premature," he says. "It can take years to figure out what can make a replication fail or succeed. You need a meta-analysis of many experiments."

Stéphane Doyen, a cognitive psychologist at the Free University of Brussels, encountered similar issues when he and his colleagues failed to replicate a classic experiment by John Bargh from Yale University in New Haven, Connecticut, showing that people walk more slowly if they have been unconsciously primed with age-related words[10]. After several rejections, Doyen's paper was also eventually published in *PLoS ONE*[11], and drew an irate blog post from Bargh. Bargh described Doyen's team as "inexpert researchers" and later took issue with the writer of this story for a blog post about the exchange. Bargh says that he responded so strongly partly because he saw growing scepticism of the idea that unconscious thought processes are important, and felt that damage was being done to the field.

Of course, one negative replication does not invalidate the original result. There are many mundane reasons why such attempts might not succeed. If the original effect is small, negative results will arise through chance alone. The volunteers in a replication attempt might differ from those in the original. And one team might simply lack the skill to reproduce another's experiments.

"The conduct of subtle experiments has much in common with the direction of a theatre performance," says Daniel Kahneman, a Nobel-prizewinning psychologist at Princeton University in New Jersey. Trivial details such as the day of the week or the colour of a room could affect the results, and these subtleties never make it into methods sections. Bargh argues, for example, that Doyen's team exposed its volunteers to too many age-related words, which could have drawn their attention to the experiment's hidden purpose. In priming studies, "you must tweak the situation just so, to make the manipulation strong enough to work, but not salient enough to attract even a little attention", says Kahneman. "Bargh has a knack that not all of us have." Kahneman says that he attributes a special 'knack' only to those who have found an effect that has been reproduced in hundreds of experiments. Bargh says of his priming experiments that he "never wanted there to be some secret knowledge about how to make these effects happen. We've always tried to give that knowledge away but maybe we should specify more details about how to do these things".

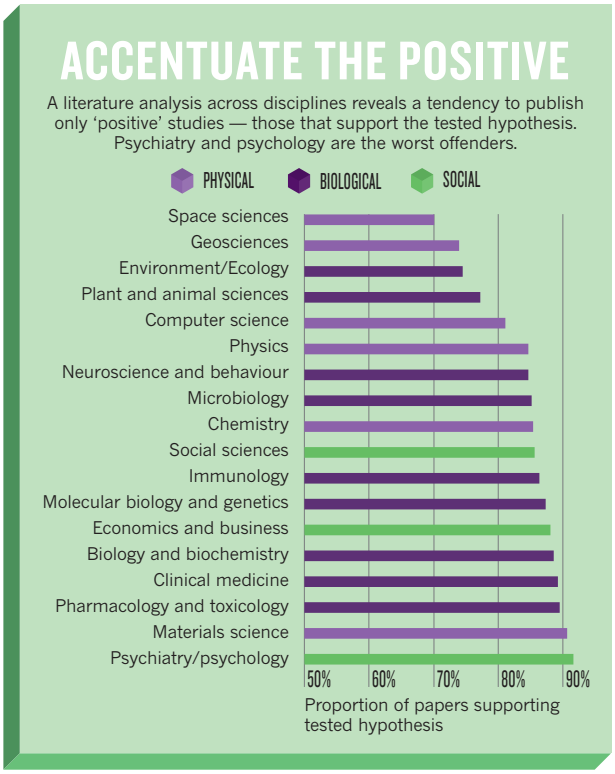
After Bargh's 1996 paper on unconscious priming, dozens of other labs followed suit with their own versions of priming experiments. Volunteers who were primed

NATURE.COM
To read a World View on lessons from fraud in psychology: go.nature.com/bkndew

by holding a heavy clipboard, for example, took interview candidates more seriously and deemed social problems to be more pressing than did those who held light boards[12]. And people primed with words relating to cleanliness judged dirty deeds more leniently[13].

Such conceptual replications are useful for psychology, which often deals with abstract concepts. "The usual way of thinking would be that [a conceptual replication] is even stronger than an exact replication. It gives better evidence for the generalizability of the effect," says Eliot Smith, a psychologist at Indiana University in Bloomington and an editor of *JPSP*.

But to other psychologists, reliance on conceptual replication is problematic. "You can't replicate a concept," says Chambers. "It's so subjective. It's anybody's guess as to how similar something needs to be to count as a conceptual replication." The practice also produces a "logical double-standard", he says. For example, if a heavy clipboard unconsciously influences people's judgements, that could be taken to conceptually replicate the slow-walking effect. But if the weight of the clipboard had no influence, no one would argue that priming had been conceptually falsified. With its ability to verify but not falsify, conceptual replication allows weak results to support one another. "It is the scientific embodiment of confirmation bias," says Brian Nosek, a social psychologist from the University of Virginia in Charlottesville. "Psychology would suffer if it wasn't practised but it doesn't replace direct replication. To show that 'A' is true, you don't do 'B'. You do 'A' again."



MISSED MISCONDUCT

These practices can create an environment in which misconduct goes undetected. In November 2011, Diederik Stapel, a social psychologist from Tilburg University in the Netherlands and a rising star in the field, was investigated for, and eventually confessed to, scientific fraud on a massive scale. Stapel had published a stream of sexy, attention-grabbing studies, showing for example that disordered environments, such as a messy train station, promote discrimination[14]. But all the factors making replication difficult helped him to cover his tracks. The scientific committee that investigated his case wrote, "Whereas all these excessively neat findings should have provoked thought, they were embraced … People accepted, if they even attempted to replicate the results for themselves, that they had failed because they lacked Mr Stapel's skill." It is now clear that Stapel manipulated and fabricated data in at least 30 publications.

Stapel's story mirrors those of psychologists Karen Ruggiero and Marc Hauser from Harvard University in Cambridge, Massachusetts, who published high-profile results on discrimination and morality, respectively. Ruggiero was found guilty of research fraud in 2001 and Hauser was found guilty of misconduct in 2010. Like Stapel, they were exposed by internal whistle-blowers. "If the field was truly self-correcting, why didn't we correct any single one of them?" asks Nosek.

Driven by these controversies, many psychologists are now searching for ways to encourage replications. "I think psychology has taken the lead in addressing this challenge," says Jonathan Schooler, a cognitive psychologist at the University of California, Santa Barbara. In January, Hal Pashler, a psychologist from the University of California, San Diego, in La Jolla and his colleagues created a website called PsychFileDrawer where psychologists can submit unpublished replication attempts, whether successful or not. The site has been warmly received but has only nine entries so far. There are few incentives to submit: any submission opens up scientists to criticisms from colleagues and does little to help their publication record.

Matthew Lieberman, a social psychologist from University of California, Los Angeles, suggests a different approach. "The top psychology programmes in the United States could require graduate students to replicate one of several nominated studies within their own field," he says. The students would build their skills and get valuable early publications, he says, and the field would learn whether surprising effects hold up.

Wagenmakers argues that replication attempts should also be published under different rules. Like clinical trials in medicine, he says, they should be pre-registered to avoid the post-hoc data-torturing practices that Simmons describes, and published irrespective of outcome. Engaging or even collaborating with the original authors early on could pre-empt any later quibbling over methods.

These changes may be a far-off hope. Some scientists still question whether there is a problem, and even Nosek points out that there are no solid estimates of the prevalence of false positives. To remedy that, late last year, he brought together a group of psychologists to try to reproduce every study published in three major psychological journals in 2008. The teams will adhere to the original experiments as closely as possible and try to work with the original authors. The goal is not to single out individual work, but to "get some initial evidence about the odds of replication" across the field, Nosek says.

Some researchers are agnostic about the outcome, but Pashler expects to see confirmation of his fears: that the corridor gossip about irreproducible studies and the file drawers stuffed with failed attempts at replication will turn out to be real. "Then, people won't be able to dodge it," he says. ■

Ed Yong *is a freelance writer based in London and author of the blog 'Not Exactly Rocket Science'.*

1. Bem, D. J. *J. Pers. Soc. Psych.* **100,** 407–425 (2011).
2. Ioannidis, J. P. A. *PLoS Med* **2,** e124 (2005).
3. Fanelli, D. *Scientometrics* **90,** 891–904 (2011).
4. Fanelli, D. *PLoS ONE* **5,** e10068 (2010).
5. Sterling, T. D. *J. Am. Stat. Assoc.* **54,** 30–34 (1959).
6. Sterling, T. D., Rosenbaum, W. L. & Weinkam, J. J. *Am. Stat.* **49,** 108–112 (1995).
7. Simmons, J. P., Nelson, L. D. & Simonsohn, U. *Psychol. Sci.* **22,** 1359–1366 (2011).
8. John, L. K., Loewenstein, G. & Prelec, D. *Psychol. Sci.* http://dx.doi.org/10.1177/0956797611430953 (2012).
9. Ritchie, S. J., Wiseman, R. & French, C. C. *PLoS ONE* **7,** e33423 (2012).
10. Bargh, J. A., Chen, M., Burrows, L. *J. Pers. Soc. Psych.* **71,** 230–244 (1996).
11. Doyen, S., Klein, O., Pichon, C.-L. & Cleeremans, A. *PLoS ONE* **7,** e29081 (2012).
12. Jostmann, N. B, Lakens, D. & Schubert, T. W. *Psychol. Sci.* **20,** 1169–1174 (2009).
13. Schnall, K, Benton, J. & Harvey, S. *Psychol. Sci.* **19,** 1219–1222 (2008).
14. Stapel, D. A. & Lindenberg, S. *Science* **332,** 251–253 (2011).