

ANNALS OF SCIENCE DECEMBER 13, 2010 ISSUE

THE TRUTH WEARS OFF

*Is there something wrong with the scientific method?***By Jonah Lehrer**

December 6, 2010

On September 18, 2007, a few dozen neuroscientists, psychiatrists, and drug-company executives gathered in a hotel conference room in Brussels to hear some startling news. It had to do with a class of drugs known as atypical or second-generation antipsychotics, which came on the market in the early nineties. The drugs, sold under brand names such as Abilify, Seroquel, and Zyprexa, had been tested on schizophrenics in several large clinical trials, all of which had demonstrated a dramatic decrease in the subjects' psychiatric symptoms. As a result, second-generation antipsychotics had become one of the fastest-growing and most profitable pharmaceutical classes. By 2001, Eli Lilly's Zyprexa was generating more revenue than Prozac. It remains the company's top-selling drug.



Many results that are rigorously proved and accepted start shrinking in later studies. Illustration by LAURENT CILLUFFO

But the data presented at the Brussels meeting made it clear that something strange was happening: the therapeutic power of the drugs appeared to be steadily waning. A recent study showed an effect that was less than half of that documented in the first trials, in the early nineteen-nineties. Many researchers began to argue that the expensive pharmaceuticals weren't any better than first-generation

antipsychotics, which have been in use since the fifties. “In fact, sometimes they now look even worse,” John Davis, a professor of psychiatry at the University of Illinois at Chicago, told me.

Before the effectiveness of a drug can be confirmed, it must be tested and tested again. Different scientists in different labs need to repeat the protocols and publish their results. The test of replicability, as it’s known, is the foundation of modern research. Replicability is how the community enforces itself. It’s a safeguard for the creep of subjectivity. Most of the time, scientists know what results they want, and that can influence the results they get. The premise of replicability is that the scientific community can correct for these flaws.

But now all sorts of well-established, multiply confirmed findings have started to look increasingly uncertain. It’s as if our facts were losing their truth: claims that have been enshrined in textbooks are suddenly unprovable. This phenomenon doesn’t yet have an official name, but it’s occurring across a wide range of fields, from psychology to ecology. In the field of medicine, the phenomenon seems extremely widespread, affecting not only antipsychotics but also therapies ranging from cardiac stents to Vitamin E and antidepressants: Davis has a forthcoming analysis demonstrating that the efficacy of antidepressants has gone down as much as threefold in recent decades.

For many scientists, the effect is especially troubling because of what it exposes about the scientific process. If replication is what separates the rigor of science from the squishiness of pseudoscience, where do we put all these rigorously validated findings that can no longer be proved? Which results should we believe? Francis Bacon, the early-modern philosopher and pioneer of the scientific method, once declared that experiments were essential, because they allowed us to “put nature to the question.” But it appears that nature often gives us different answers.

Jonathan Schooler was a young graduate student at the University of Washington in the nineteen-eighties when he discovered a surprising new fact about language and memory. At the time, it was widely believed that the act of describing our memories improved them. But, in a series of clever experiments, Schooler demonstrated that subjects shown a face and asked to describe it were much less likely to recognize the face when shown it later than those who had simply looked at it. Schooler called the phenomenon “verbal overshadowing.”

The study turned him into an academic star. Since its initial publication, in 1990, it has been cited more than four hundred times. Before long, Schooler had extended the model to a variety of other tasks, such as remembering the taste of a wine, identifying the best strawberry jam, and solving difficult creative puzzles. In each instance, asking people to put their perceptions into words led to dramatic decreases in performance.

But while Schooler was publishing these results in highly reputable journals, a secret worry gnawed at him: it was proving difficult to replicate his earlier findings. “I’d often still see an effect, but the effect just wouldn’t be as strong,” he told me. “It was as if verbal overshadowing, my big new idea, was getting weaker.” At first, he assumed that he’d made an error in experimental design or a statistical miscalculation. But he couldn’t find anything wrong with his research. He then concluded that his initial batch of research subjects must have been unusually susceptible to verbal overshadowing. (John Davis, similarly, has speculated that part of the drop-off in the effectiveness of antipsychotics can be attributed to using subjects who suffer from milder forms of psychosis which are less likely to show dramatic improvement.) “It wasn’t a very satisfying explanation,” Schooler says. “One of my mentors told me that my real mistake was trying to replicate my work. He told me doing that was just setting myself up for disappointment.”

Schooler tried to put the problem out of his mind; his colleagues assured him that such things happened all the time. Over the next few years, he found new research questions, got married and had kids. But his replication problem kept on getting worse. His first attempt at replicating the 1990 study, in 1995, resulted in an effect that was thirty per cent smaller. The next year, the size of the effect shrank another thirty per cent. When other labs repeated Schooler’s experiments, they got a similar spread of data, with a distinct downward trend. “This was profoundly frustrating,” he says. “It was as if nature gave me this great result and then tried to take it back.” In private, Schooler began referring to the problem as “cosmic habituation,” by analogy to the decrease in response that occurs when individuals habituate to particular stimuli. “Habituation is why you don’t notice the stuff that’s always there,” Schooler says. “It’s an inevitable process of adjustment, a ratcheting down of excitement. I started joking that it was like the cosmos was habituating to my ideas. I took it very personally.”

Schooler is now a tenured professor at the University of California at Santa Barbara. He has curly black hair, pale-green eyes, and the relaxed demeanor of someone who lives five minutes away from

his favorite beach. When he speaks, he tends to get distracted by his own digressions. He might begin with a point about memory, which reminds him of a favorite William James quote, which inspires a long soliloquy on the importance of introspection. Before long, we're looking at pictures from Burning Man on his iPhone, which leads us back to the fragile nature of memory.

Although verbal overshadowing remains a widely accepted theory—it's often invoked in the context of eyewitness testimony, for instance—Schooler is still a little peeved at the cosmos. "I know I should just move on already," he says. "I really should stop talking about this. But I can't." That's because he is convinced that he has stumbled on a serious problem, one that afflicts many of the most exciting new ideas in psychology.

One of the first demonstrations of this mysterious phenomenon came in the early nineteen-thirties. Joseph Banks Rhine, a psychologist at Duke, had developed an interest in the possibility of extrasensory perception, or E.S.P. Rhine devised an experiment featuring Zener cards, a special deck of twenty-five cards printed with one of five different symbols: a card was drawn from the deck and the subject was asked to guess the symbol. Most of Rhine's subjects guessed about twenty per cent of the cards correctly, as you'd expect, but an undergraduate named Adam Linzmayer averaged nearly fifty per cent during his initial sessions, and pulled off several uncanny streaks, such as guessing nine cards in a row. The odds of this happening by chance are about one in two million. Linzmayer did it three times.

Rhine documented these stunning results in his notebook and prepared several papers for publication. But then, just as he began to believe in the possibility of extrasensory perception, the student lost his spooky talent. Between 1931 and 1933, Linzmayer guessed at the identity of another several thousand cards, but his success rate was now barely above chance. Rhine was forced to conclude that the student's "extra-sensory perception ability has gone through a marked decline." And Linzmayer wasn't the only subject to experience such a drop-off: in nearly every case in which Rhine and others documented E.S.P. the effect dramatically diminished over time. Rhine called this trend the "decline effect."

Schooler was fascinated by Rhine's experimental struggles. Here was a scientist who had repeatedly documented the decline of his data; he seemed to have a talent for finding results that fell apart. In 2004, Schooler embarked on an ironic imitation of Rhine's research: he tried to replicate this failure

to replicate. In homage to Rhine's interests, he decided to test for a parapsychological phenomenon known as precognition. The experiment itself was straightforward: he flashed a set of images to a subject and asked him or her to identify each one. Most of the time, the response was negative—the images were displayed too quickly to register. Then Schooler randomly selected half of the images to be shown again. What he wanted to know was whether the images that got a second showing were more likely to have been identified the first time around. Could subsequent exposure have somehow influenced the initial results? Could the effect become the cause?

The craziness of the hypothesis was the point: Schooler knows that precognition lacks a scientific explanation. But he wasn't testing extrasensory powers; he was testing the decline effect. "At first, the data looked amazing, just as we'd expected," Schooler says. "I couldn't believe the amount of precognition we were finding. But then, as we kept on running subjects, the effect size"—a standard statistical measure—"kept on getting smaller and smaller." The scientists eventually tested more than two thousand undergraduates. "In the end, our results looked just like Rhine's," Schooler said. "We found this strong paranormal effect, but it disappeared on us."

The most likely explanation for the decline is an obvious one: regression to the mean. As the experiment is repeated, that is, an early statistical fluke gets cancelled out. The extrasensory powers of Schooler's subjects didn't decline—they were simply an illusion that vanished over time. And yet Schooler has noticed that many of the data sets that end up declining seem statistically solid—that is, they contain enough data that any regression to the mean shouldn't be dramatic. "These are the results that pass all the tests," he says. "The odds of them being random are typically quite remote, like one in a million. This means that the decline effect should almost never happen. But it happens all the time! Hell, it's happened to me multiple times." And this is why Schooler believes that the decline effect deserves more attention: its ubiquity seems to violate the laws of statistics. "Whenever I start talking about this, scientists get very nervous," he says. "But I still want to know what happened to my results. Like most scientists, I assumed that it would get easier to document my effect over time. I'd get better at doing the experiments, at zeroing in on the conditions that produce verbal overshadowing. So why did the opposite happen? I'm convinced that we can use the tools of science to figure this out. First, though, we have to admit that we've got a problem."

In 1991, the Danish zoologist Anders Møller, at Uppsala University, in Sweden, made a remarkable discovery about sex, barn swallows, and symmetry. It had long been known that the asymmetrical

Appearance of a creature was directly linked to the amount of mutation in its genome, so that more mutations led to more “fluctuating asymmetry.” (An easy way to measure asymmetry in humans is to compare the length of the fingers on each hand.) What Møller discovered is that female barn swallows were far more likely to mate with male birds that had long, symmetrical feathers. This suggested that the picky females were using symmetry as a proxy for the quality of male genes. Møller’s paper, which was published in *Nature*, set off a frenzy of research. Here was an easily measured, widely applicable indicator of genetic quality, and females could be shown to gravitate toward it. Aesthetics was really about genetics.

In the three years following, there were ten independent tests of the role of fluctuating asymmetry in sexual selection, and nine of them found a relationship between symmetry and male reproductive success. It didn’t matter if scientists were looking at the hairs on fruit flies or replicating the swallow studies—females seemed to prefer males with mirrored halves. Before long, the theory was applied to humans. Researchers found, for instance, that women preferred the smell of symmetrical men, but only during the fertile phase of the menstrual cycle. Other studies claimed that females had more orgasms when their partners were symmetrical, while a paper by anthropologists at Rutgers analyzed forty Jamaican dance routines and discovered that symmetrical men were consistently rated as better dancers.

Then the theory started to fall apart. In 1994, there were fourteen published tests of symmetry and sexual selection, and only eight found a correlation. In 1995, there were eight papers on the subject, and only four got a positive result. By 1998, when there were twelve additional investigations of fluctuating asymmetry, only a third of them confirmed the theory. Worse still, even the studies that yielded some positive result showed a steadily declining effect size. Between 1992 and 1997, the average effect size shrank by eighty per cent.

And it’s not just fluctuating asymmetry. In 2001, Michael Jennions, a biologist at the Australian National University, set out to analyze “temporal trends” across a wide range of subjects in ecology and evolutionary biology. He looked at hundreds of papers and forty-four meta-analyses (that is, statistical syntheses of related studies), and discovered a consistent decline effect over time, as many of the theories seemed to fade into irrelevance. In fact, even when numerous variables were controlled for—Jennions knew, for instance, that the same author might publish several critical papers, which could distort his analysis—there was still a significant decrease in the validity of the hypothesis, often

within a year of publication. Jennions admits that his findings are troubling, but expresses a reluctance to talk about them publicly. “This is a very sensitive issue for scientists,” he says. “You know, we’re supposed to be dealing with hard facts, the stuff that’s supposed to stand the test of time. But when you see these trends you become a little more skeptical of things.”

What happened? Leigh Simmons, a biologist at the University of Western Australia, suggested one explanation when he told me about his initial enthusiasm for the theory: “I was really excited by fluctuating asymmetry. The early studies made the effect look very robust.” He decided to conduct a few experiments of his own, investigating symmetry in male horned beetles. “Unfortunately, I couldn’t find the effect,” he said. “But the worst part was that when I submitted these null results I had difficulty getting them published. The journals only wanted confirming data. It was too exciting an idea to disprove, at least back then.” For Simmons, the steep rise and slow fall of fluctuating asymmetry is a clear example of a scientific paradigm, one of those intellectual fads that both guide and constrain research: after a new paradigm is proposed, the peer-review process is tilted toward positive results. But then, after a few years, the academic incentives shift—the paradigm has become entrenched—so that the most notable results are now those that disprove the theory.

Jennions, similarly, argues that the decline effect is largely a product of publication bias, or the tendency of scientists and scientific journals to prefer positive data over null results, which is what happens when no effect is found. The bias was first identified by the statistician Theodore Sterling, in 1959, after he noticed that ninety-seven per cent of all published psychological studies with statistically significant data found the effect they were looking for. A “significant” result is defined as any data point that would be produced by chance less than five per cent of the time. This ubiquitous test was invented in 1922 by the English mathematician Ronald Fisher, who picked five per cent as the boundary line, somewhat arbitrarily, because it made pencil and slide-rule calculations easier. Sterling saw that if ninety-seven per cent of psychology studies were proving their hypotheses, either psychologists were extraordinarily lucky or they published only the outcomes of successful experiments. In recent years, publication bias has mostly been seen as a problem for clinical trials, since pharmaceutical companies are less interested in publishing results that aren’t favorable. But it’s becoming increasingly clear that publication bias also produces major distortions in fields without large corporate incentives, such as psychology and ecology.

hile publication bias almost certainly plays a role in the decline effect, it remains an incomplete

W explanation. For one thing, it fails to account for the initial prevalence of positive results among studies that never even get submitted to journals. It also fails to explain the experience of people like Schooler, who have been unable to replicate their initial data despite their best efforts. Richard Palmer, a biologist at the University of Alberta, who has studied the problems surrounding fluctuating asymmetry, suspects that an equally significant issue is the selective reporting of results—the data that scientists choose to document in the first place. Palmer’s most convincing evidence relies on a statistical tool known as a funnel graph. When a large number of studies have been done on a single subject, the data should follow a pattern: studies with a large sample size should all cluster around a common value—the true result—whereas those with a smaller sample size should exhibit a random scattering, since they’re subject to greater sampling error. This pattern gives the graph its name, since the distribution resembles a funnel.

The funnel graph visually captures the distortions of selective reporting. For instance, after Palmer plotted every study of fluctuating asymmetry, he noticed that the distribution of results with smaller sample sizes wasn’t random at all but instead skewed heavily toward positive results. Palmer has since documented a similar problem in several other contested subject areas. “Once I realized that selective reporting is everywhere in science, I got quite depressed,” Palmer told me. “As a researcher, you’re always aware that there might be some nonrandom patterns, but I had no idea how widespread it is.” In a recent review article, Palmer summarized the impact of selective reporting on his field: “We cannot escape the troubling conclusion that some—perhaps many—cherished generalities are at best exaggerated in their biological significance and at worst a collective illusion nurtured by strong a-priori beliefs often repeated.”

Palmer emphasizes that selective reporting is not the same as scientific fraud. Rather, the problem seems to be one of subtle omissions and unconscious misperceptions, as researchers struggle to make sense of their results. Stephen Jay Gould referred to this as the “shoehorning” process. “A lot of scientific measurement is really hard,” Simmons told me. “If you’re talking about fluctuating asymmetry, then it’s a matter of minuscule differences between the right and left sides of an animal. It’s millimetres of a tail feather. And so maybe a researcher knows that he’s measuring a good male”—an animal that has successfully mated—“and he knows that it’s supposed to be symmetrical. Well,

that act of measurement is going to be vulnerable to all sorts of perception biases. That's not a cynical statement. That's just the way human beings work."

One of the classic examples of selective reporting concerns the testing of acupuncture in different countries. While acupuncture is widely accepted as a medical treatment in various Asian countries, its use is much more contested in the West. These cultural differences have profoundly influenced the results of clinical trials. Between 1966 and 1995, there were forty-seven studies of acupuncture in China, Taiwan, and Japan, and every single trial concluded that acupuncture was an effective treatment. During the same period, there were ninety-four clinical trials of acupuncture in the United States, Sweden, and the U.K., and only fifty-six per cent of these studies found any therapeutic benefits. As Palmer notes, this wide discrepancy suggests that scientists find ways to confirm their preferred hypothesis, disregarding what they don't want to see. Our beliefs are a form of blindness.

John Ioannidis, an epidemiologist at Stanford University, argues that such distortions are a serious issue in biomedical research. "These exaggerations are why the decline has become so common," he says. "It'd be really great if the initial studies gave us an accurate summary of things. But they don't. And so what happens is we waste a lot of money treating millions of patients and doing lots of follow-up studies on other themes based on results that are misleading." In 2005, Ioannidis published an article in the *Journal of the American Medical Association* that looked at the forty-nine most cited clinical-research studies in three major medical journals. Forty-five of these studies reported positive results, suggesting that the intervention being tested was effective. Because most of these studies were randomized controlled trials—the "gold standard" of medical evidence—they tended to have a significant impact on clinical practice, and led to the spread of treatments such as hormone replacement therapy for menopausal women and daily low-dose aspirin to prevent heart attacks and strokes. Nevertheless, the data Ioannidis found were disturbing: of the thirty-four claims that had been subject to replication, forty-one per cent had either been directly contradicted or had their effect sizes significantly downgraded.

The situation is even worse when a subject is fashionable. In recent years, for instance, there have been hundreds of studies on the various genes that control the differences in disease risk between men and women. These findings have included everything from the mutations responsible for the increased risk of schizophrenia to the genes underlying hypertension. Ioannidis and his colleagues looked at four hundred and thirty-two of these claims. They quickly discovered that the vast majority

had serious flaws. But the most troubling fact emerged when he looked at the test of replication: out of four hundred and thirty-two claims, only a single one was consistently replicable. “This doesn’t mean that none of these claims will turn out to be true,” he says. “But, given that most of them were done badly, I wouldn’t hold my breath.”

According to Ioannidis, the main problem is that too many researchers engage in what he calls “significance chasing,” or finding ways to interpret the data so that it passes the statistical test of significance—the ninety-five-per-cent boundary invented by Ronald Fisher. “The scientists are so eager to pass this magical test that they start playing around with the numbers, trying to find anything that seems worthy,” Ioannidis says. In recent years, Ioannidis has become increasingly blunt about the pervasiveness of the problem. One of his most cited papers has a deliberately provocative title: “Why Most Published Research Findings Are False.”

The problem of selective reporting is rooted in a fundamental cognitive flaw, which is that we like proving ourselves right and hate being wrong. “It feels good to validate a hypothesis,” Ioannidis said. “It feels even better when you’ve got a financial interest in the idea or your career depends upon it. And that’s why, even after a claim has been systematically disproven”—he cites, for instance, the early work on hormone replacement therapy, or claims involving various vitamins—“you still see some stubborn researchers citing the first few studies that show a strong effect. They really want to believe that it’s true.”

That’s why Schooler argues that scientists need to become more rigorous about data collection before they publish. “We’re wasting too much time chasing after bad studies and underpowered experiments,” he says. The current “obsession” with replicability distracts from the real problem, which is faulty design. He notes that nobody even tries to replicate most science papers—there are simply too many. (According to *Nature*, a third of all studies never even get cited, let alone repeated.) “I’ve learned the hard way to be exceedingly careful,” Schooler says. “Every researcher should have to spell out, in advance, how many subjects they’re going to use, and what exactly they’re testing, and what constitutes a sufficient level of proof. We have the tools to be much more transparent about our experiments.”

In a forthcoming paper, Schooler recommends the establishment of an open-source database, in which researchers are required to outline their planned investigations and document all their results.

“I think this would provide a huge increase in access to scientific work and give us a much better way to judge the quality of an experiment,” Schooler says. “It would help us finally deal with all these issues that the decline effect is exposing.”

Although such reforms would mitigate the dangers of publication bias and selective reporting, they still wouldn’t erase the decline effect. This is largely because scientific research will always be shadowed by a force that can’t be curbed, only contained: sheer randomness. Although little research has been done on the experimental dangers of chance and happenstance, the research that exists isn’t encouraging.

In the late nineteen-nineties, John Crabbe, a neuroscientist at the Oregon Health and Science University, conducted an experiment that showed how unknowable chance events can skew tests of replicability. He performed a series of experiments on mouse behavior in three different science labs: in Albany, New York; Edmonton, Alberta; and Portland, Oregon. Before he conducted the experiments, he tried to standardize every variable he could think of. The same strains of mice were used in each lab, shipped on the same day from the same supplier. The animals were raised in the same kind of enclosure, with the same brand of sawdust bedding. They had been exposed to the same amount of incandescent light, were living with the same number of littermates, and were fed the exact same type of chow pellets. When the mice were handled, it was with the same kind of surgical glove, and when they were tested it was on the same equipment, at the same time in the morning.

The premise of this test of replicability, of course, is that each of the labs should have generated the same pattern of results. “If any set of experiments should have passed the test, it should have been ours,” Crabbe says. “But that’s not the way it turned out.” In one experiment, Crabbe injected a particular strain of mouse with cocaine. In Portland the mice given the drug moved, on average, six hundred centimetres more than they normally did; in Albany they moved seven hundred and one additional centimetres. But in the Edmonton lab they moved more than five thousand additional centimetres. Similar deviations were observed in a test of anxiety. Furthermore, these inconsistencies didn’t follow any detectable pattern. In Portland one strain of mouse proved most anxious, while in Albany another strain won that distinction.

The disturbing implication of the Crabbe study is that a lot of extraordinary scientific data are nothing but noise. The hyperactivity of those coked-up Edmonton mice wasn’t an interesting new

fact—it was a meaningless outlier, a by-product of invisible variables we don't understand. The problem, of course, is that such dramatic findings are also the most likely to get published in prestigious journals, since the data are both statistically significant and entirely unexpected. Grants get written, follow-up studies are conducted. The end result is a scientific accident that can take years to unravel.

This suggests that the decline effect is actually a decline of illusion. While Karl Popper imagined falsification occurring with a single, definitive experiment—Galileo refuted Aristotelian mechanics in an afternoon—the process turns out to be much messier than that. Many scientific theories continue to be considered true even after failing numerous experimental tests. Verbal overshadowing might exhibit the decline effect, but it remains extensively relied upon within the field. The same holds for any number of phenomena, from the disappearing benefits of second-generation antipsychotics to the weak coupling ratio exhibited by decaying neutrons, which appears to have fallen by more than ten standard deviations between 1969 and 2001. Even the law of gravity hasn't always been perfect at predicting real-world phenomena. (In one test, physicists measuring gravity by means of deep boreholes in the Nevada desert found a two-and-a-half-per-cent discrepancy between the theoretical predictions and the actual data.) Despite these findings, second-generation antipsychotics are still widely prescribed, and our model of the neutron hasn't changed. The law of gravity remains the same.

Such anomalies demonstrate the slipperiness of empiricism. Although many scientific ideas generate conflicting results and suffer from falling effect sizes, they continue to get cited in the textbooks and drive standard medical practice. Why? Because these ideas seem true. Because they make sense. Because we can't bear to let them go. And this is why the decline effect is so troubling. Not because it reveals the human fallibility of science, in which data are tweaked and beliefs shape perceptions. (Such shortcomings aren't surprising, at least for scientists.) And not because it reveals that many of our most exciting theories are fleeting fads and will soon be rejected. (That idea has been around since Thomas Kuhn.) The decline effect is troubling because it reminds us how difficult it is to prove anything. We like to pretend that our experiments define the truth for us. But that's often not the case. Just because an idea is true doesn't mean it can be proved. And just because an idea can be proved doesn't mean it's true. When the experiments are done, we still have to choose what to believe.



Published in the print edition of the December 13, 2010 issue.