THE UNSCIENTIFIC METHOD

alarming amount of
earch is flawed because
unconscious biases.
iat's to be done, asks
nia van Gilder Cooke

LISTENING to *When I'm Sixty-Four* by The Beatles can make you younger. This miraculous effect, dubbed "chronological rejuvenation", was revealed in the journal *Psychological Science* in 2011. It wasn't a hoax, but you'd be right to be suspicious. The aim was to show how easy it is to generate statistical evidence for pretty much anything, simply by picking and choosing methods and data in ways that researchers do every day.

The paper caused a stir among psychologists, and has become the most cited in the journal's history. The following year, Nobel prizewinning psychologist Daniel Kahneman stoked the fire with an open email to social psychologists warning of a "train wreck" if they didn't clean up their act. But things only came to a head last year with the publication of a paper in *Science*. It described a major effort to replicate 100 psychology experiments published in top journals. The success rate was little more than a third. People began to talk of a "crisis" in psychology.

In fact, the problem extends far beyond psychology – dubious results are alarmingly common in many fields of science. Worryingly, they seem to be especially shaky in areas that have a direct bearing on human well-being – the science underpinning everyday political, economic and healthcare decisions. No wonder the whistle-blowers are urgently trying to investigate why it's happening, how big the problem is and what can be done to fix it. In doing so, they are highlighting flaws in the way we all think, and exposing cracks in the culture of science.

Science is often thought of as a dispassionate search for the truth. But, of course, we are all only human. And most people want to climb the professional ladder. The main way to do that if you're a scientist is to get grants and publish lots of papers. The problem is that journals have a clear preference for research showing strong, positive relationships – between a particular medical treatment and improved health, for example. This means researchers often try to find those sorts of results. A few go as far as making things up. But a huge number tinker with their research in ways they think are harmless, but which can bias the outcome.

This tinkering can take many forms (see "To err is human", page 40). You peek at the results and stop an experiment when it shows what you were expecting. You throw out data points that don't fit your hypothesis – something could be wrong with those results, you reason. Or you run several types of statistical analysis and end up using the one that shows the strongest effect. "It can be very hard to even see that biases might be entering your reasoning," says psychologist Brian Nosek at the University of Virginia in Charlottesville, who led the team trying to replicate 100 psychology studies. Take the tendency to scrutinise results that don't fit with your predictions more carefully than those that do. "There's no nefarious motive," says Roger Peng at Johns Hopkins University in Baltimore, Maryland. It's just natural to assume these results are likely to be "wrong".

You might think that journals, which get peers from the same scientific field to review papers, would pick up on such practices. But, say critics, the system isn't up to the task. For one thing, most journals don't ask researchers to give them a tour of their statistical sausage factory. "The vast majority don't require that you make any data available beyond a brief description of the methods," says Peng. Peer-reviewers usually don't see the complete data and methods either. And even if they did, they might not have the time, ability or inclination to check them. Refereeing is unpaid and ›

TO ERR IS HUMAN

Bias is inherent in research but there are ways to limit it

PROBLEMS:

Wishful thinking – Unconsciously biasing methods to confirm your hypothesis

Sneaky stats – Using the statistical analysis that best supports your hypothesis

Burying evidence – Not sharing research data so that results can be scrutinised

Rewriting history – Inventing a new hypothesis to explain unexpected results

Tidying up – Ignoring inconvenient data points and analyses in the write-up

FIXES:

Pre-registration – Publicly declaring procedures before doing a study

Blindfolding – Deciding on a data analysis method before the data are collected

Sharing – Making methods and data transparent and available to others

Collaboration – Working with others to increase the rigour of experiments

Statistical education – Acquiring the tools required to assess data meaningfully

"In a major effort to replicate 100 psychology experiments the success rate was little more than a third"

anonymous – so there's no reward and no recognition in it.

All this helps explain why so many studies don't hold up when others try to replicate them. But it doesn't explain why psychology in particular is facing a "crisis" right now. There's nothing new about researchers being subconsciously committed to proving their own theories, or journals favouring headline-grabbing research. Sure, the pressure on researchers to publish is ever greater, however, what's really new is the scrutiny being given to their published findings.

Traditionally, once results are published they tend to go unchecked. "The current system does not reward replication – it often even penalizes people who want to rigorously replicate previous work," wrote statistician John Ioannidis of Stanford University in California in a recent paper entitled "How to make more published research true". Proponents of a new discipline called metascience (the science of science) aim to change that, and Ioannidis is in the vanguard.

Psychology may have borne the brunt of the controversy so far, but Ioannidis has for a long time argued that the problem is widespread. In 2005, he claimed that sloppy methods could mean more than half of all published scientific results are flawed. Some fields of research are less susceptible than others, though. In astronomy, chemistry and physics, for instance, "people have a very strong tradition of sharing data, and of using common databases like big telescopes or high energy physical experiments", Ioannidis says. "They are very cautious about making claims that eventually will be refuted." But in fields where such checks and balances are absent, irreproducible results are rife.

Take the case of cancer researcher Anil Potti when he was at Duke University in Durham, North Carolina. In 2006, staff at the MD Anderson Cancer Center in Houston, Texas, wanted to investigate treatments based on Potti's published work on gene expression. Before pressing ahead, they asked their colleagues, biostatisticians Keith Baggerly and Kevin Coombes, to look over the findings. Their efforts illustrate how hard it can be for peer reviewers to pick up on mistakes. It took them almost 2000 hours to disentangle the data and reveal a catalogue of errors. It later transpired that Potti had falsified data, but in the meantime, three clinical trials had been started on the basis of his research.

Evidence is mounting that medical research is particularly prone to irreproducibility. In 2012, Glenn Begley, a biotech consultant,

Doesn't add up: unpopular austerity measures were based on imperfect maths

showed that just 11 per cent of the preclinical cancer studies coming out of the academic pipeline that he sampled were replicable. Another study estimates that irreproducible preclinical research costs the US $28 billion a year and slows down the development of life-saving drugs. "The truth is everyone knew that this was a problem," says Begley. "No one really knew the magnitude of the problem."

Dodgy statistics

It's the tip of the iceberg. Research published last year by Megan Head of the Australian National University in Canberra and her colleagues showed that dodgy statistics are rife in the biological sciences. They scrutinised results from a wide range of scientific disciplines for evidence for "p-hacking" – collecting or selecting data or statistical analyses until non-significant results becomes significant. They found it to be particularly common in biological sciences. "A lot of biologists go into biology because they don't want to do maths, and then they get a rude shock when they learn they have to do statistics," says Head.

But even mathematicians make errors. In 2010, economists Carmen Reinhart and Kenneth Rogoff at Harvard University published research showing that when a country's debt reaches more than 90 per cent of GDP there is an associated plunge in economic growth. The paper, which appeared in a non-peer-reviewed edition of the *American Economic Review*, was seized on by

STUART FRANKLIN/MAGNUM PHOTOS

"Sloppy methods could mean that over half of all published scientific results are flawed"

icians in the UK and US to justify
erity policies. However, three years later,
n Thomas Herndon, a graduate student at
Jniversity of Massachusetts, Amherst,
to replicate the findings, he ran into
ole. Reinhart and Rogoff had made several
akes including a coding error in their
adsheet. The effect they had observed had,
rding to critics, been largely a mirage.
rtheless, it had a major impact on the
ic policy debate.
ven how influential a flawed paper can be,
o wonder people are up in arms. One
ous concern is that it could undermine
ic faith in science itself. "It could very
ly become a wave of mistrust of the kind
nd associated with climate change," says
nologist Nicholas Humphrey at the
on School of Economics. Drawing an
gy with the global financial collapse of
, he calls it "sub-prime science". "After the disgrace of the bankers, science must not be next," he wrote, earlier this year.

So what can be done? There has already been a rapid response in one area of research where irreproducible results can have life-or-death consequences. Since 2005, a group of major medical journals has required researchers to publicly register clinical trials, and the methods they intend to use, before recruiting patients. Ioannidis estimates that about half of all clinical trials now are pre-registered, vastly reducing the possibility of flawed work.

Psychologists have also taken matters into their own hands. In 2011, the authors of the *When I'm Sixty-Four* paper – Joseph Simmons and Uri Simonsohn of the University of Pennsylvania and Leif Nelson of the University of California, Berkeley – met with Eric Eich, the newly appointed editor of *Psychological Science*, to discuss the problems facing their discipline. "That was really eye-opening for me," says Eich. "There were a lot of things that were essentially broken."

In January 2014, the journal began asking researchers more questions about their methods and giving them more space to explain them. It also introduced a "nudge" to reward good practice by displaying badges on papers to recognise those who made data and methods available or pre-registered their study. The result? Submissions fell off a cliff. "I thought I had broken the damn journal," says Eich. However, after five months, submission rates were back to normal, and now some 40 per cent of new *Psychological Science* papers have open data – up from 3 per cent before the badges were introduced.

Now the idea is being rolled out. Last year, Nosek and his colleagues came up with guidelines that journals could follow to increase transparency and reproducibility. These have since been endorsed by the US National Institutes of Health, and adopted by more than 500 journals, including *Science*, and 50 organisations. *Nature* has its own guidelines. Meanwhile, the Center for Open Science, co-founded by Nosek, has established a free online platform, the Open Science Framework, where researchers can register studies and display all their data and methods. More radically, there have been calls to replace peer reviewers with paid experts – accredited specialists in the analysis of research.

Quality not quantity

Universities may join the movement too. Ioannidis and others are working to create a "coalition of university leaders" to address the problem. "Universities are the gatekeepers of promotion and tenure," he says. "I hope that we will be moving pretty soon on that front." One obvious solution is to stop rewarding scientists on the basis of how much they have published – to consider quality not quantity when making academic promotions.

Ultimately, we may need to create novel ways of determining which studies are valid. Working with Nosek's team, the Science Prediction Market Project asked psychologists to place bets on which studies would stand up and which wouldn't. "It turned out that the market performed pretty well in predicting the outcome of the replications," says Anna Dreber Almenberg at the Stockholm School of Economics in Sweden, who leads the project. Such an approach could be harnessed to help identify iffy results before they are accepted for publication. It's still early days, but Dreber Almenberg says that prediction markets "could be interesting to think more about".

Meanwhile, replication projects are gaining popularity. Groups are now looking at cancer research and experimental economics. One member of the economics group, Colin Camerer at the California Institute of Technology in Pasadena, says the project, which published results of a pilot study in March, has been greeted with enthusiasm. "People have been emailing us saying, if you do more, we'll help you out," he says.

"It will take years to play out," says Eich. "But hopefully at the end of it, you get more replicable, high-quality science." Given that we fund academic research through our taxes and rely on it to improve our lives, that will be good for everybody. ■

Life-saver: new drugs are slower to emerge when cash is blown on research that can't be replicated

Sonia van Gilder Cooke is based in London