Unreliable research

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

Oct 19th 2013 | From the print edition

"I SEE a train wreck looming," warned Daniel Kahneman, an eminent psychologist, in an open letter last year. The premonition concerned research on a phenomenon known as "priming". Priming studies suggest that decisions can be influenced by apparently irrelevant actions or events that took place just before the cusp of choice. They have been a boom area in psychology over the past decade, and some of their insights have already made it out of the lab and into the toolkits of policy wonks keen on "nudging" the populace.

Dr Kahneman and a growing number of his colleagues fear that a lot of this priming research is poorly founded. Over the past few years various researchers have made systematic attempts to replicate some of the more widely cited priming experiments. Many of these replications have failed. In April, for instance, a paper in *PLoS ONE*, a journal, reported that nine separate experiments had not managed to reproduce the results of a famous study from 1998 purporting to show that thinking about a professor before taking an intelligence test leads to a higher score than imagining a football hooligan.

The idea that the same experiments always get the same results, no matter who performs them, is one of the cornerstones of science's claim to objective truth. If a systematic campaign of replication does not lead to the same results, then either the original research is flawed (as the replicators claim) or the replications are (as many of the original researchers on priming contend). Either way, something is awry.

To err is all too common

It is tempting to see the priming fracas as an isolated case in an area of science—psychology—easily marginalised as soft and wayward. But irreproducibility is much more widespread. A few years ago scientists at Amgen, an American drug company, tried to replicate 53 studies that they considered landmarks in the basic science of cancer, often co-operating closely with the original researchers to ensure that their experimental technique matched the one used first time round. According to a piece they wrote last year in *Nature*, a leading scientific journal, they were able to reproduce the original results in just six. Months earlier Florian Prinz and his colleagues at Bayer HealthCare, a German pharmaceutical giant, reported in *Nature Reviews Drug Discovery*, a sister journal, that they had

successfully reproduced the published results in just a quarter of 67 seminal studies.

The governments of the OECD, a club of mostly rich countries, spent $59 billion on biomedical research in 2012, nearly double the figure in 2000. One of the justifications for this is that basic-science results provided by governments form the basis for private drug-development work. If companies cannot rely on academic research, that reasoning breaks down. When an official at America's National Institutes of Health (NIH) reckons, despairingly, that researchers would find it hard to reproduce at least three-quarters of all published biomedical findings, the public part of the process seems to have failed.

Academic scientists readily acknowledge that they often get things wrong. But they also hold fast to the idea that these errors get corrected over time as other scientists try to take the work further. Evidence that many more dodgy results are published than are subsequently corrected or withdrawn calls that much-vaunted capacity for self-correction into question. There are errors in a lot more of the scientific papers being published, written about and acted on than anyone would normally suppose, or like to think.

Various factors contribute to the problem. Statistical mistakes are widespread. The peer reviewers who evaluate papers before journals commit to publishing them are much worse at spotting mistakes than they or others appreciate. Professional pressure, competition and ambition push scientists to publish more quickly than would be wise. A career structure which lays great stress on publishing copious papers exacerbates all these problems. "There is no cost to getting things wrong," says Brian Nosek, a psychologist at the University of Virginia who has taken an interest in his discipline's persistent errors. "The cost is not getting them published."

First, the statistics, which if perhaps off-putting are quite crucial. Scientists divide errors into two classes. A type I error is the mistake of thinking something is true when it is not (also known as a "false positive"). A type II error is thinking something is not true when in fact it is (a "false negative"). When testing a specific hypothesis, scientists run statistical checks to work out how likely it would be for data which seem to support the idea to have come about simply by chance. If the likelihood of such a false-positive conclusion is less than 5%, they deem the evidence that the hypothesis is true "statistically significant". They are thus accepting that one result in 20 will be falsely positive—but one in 20 seems a satisfactorily low rate.

Understanding insignificance

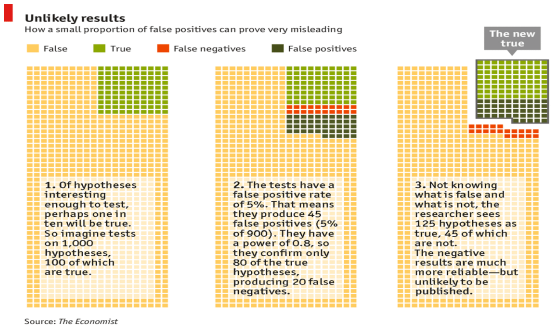
In 2005 John Ioannidis, an epidemiologist from Stanford University, caused a stir with a paper showing why, as a matter of statistical logic, the idea that only one such paper in 20 gives a false-positive result was hugely optimistic. Instead, he argued, "most published research findings are probably false." As he told the quadrennial International Congress on Peer Review and Biomedical Publication, held this September in Chicago, the problem has not gone away.

Dr Ioannidis draws his stark conclusion on the basis that the customary approach to statistical significance ignores three things: the "statistical power" of the study (a measure of its ability to avoid type II errors, false negatives in which a real signal is missed in the noise); the unlikeliness of the hypothesis being tested; and the pervasive bias favouring the publication of claims to have found something new.

A statistically powerful study is one able to pick things up even when their effects on the data are small. In general bigger studies—those which run the experiment more times, recruit more patients for the trial, or whatever—are more powerful. A power of 0.8 means that of ten true hypotheses tested, only two will be ruled out because their effects are not picked up in the data; this is widely accepted as powerful enough for most purposes. But this benchmark is not always met, not least because big studies are more expensive. A study in April by Dr Ioannidis and colleagues found that in neuroscience the typical statistical power is a dismal 0.21; writing in *Perspectives on Psychological Science*, Marjan Bakker of the University of Amsterdam and colleagues reckon that in that field the average power is 0.35.

Unlikeliness is a measure of how surprising the result might be. By and large, scientists want surprising results, and so they test hypotheses that are normally pretty unlikely and often very unlikely. Dr Ioannidis argues that in his field, epidemiology, you might expect one in ten hypotheses to be true. In exploratory disciplines like genomics, which rely on combing through vast troves of data about genes and proteins for interesting relationships, you might expect just one in a thousand to prove correct.

With this in mind, consider 1,000 hypotheses being tested of which just 100 are true (see chart). Studies with a power of 0.8 will find 80 of them, missing 20 because of false negatives. Of the 900 hypotheses that are wrong, 5%—that is, 45 of them—will look right because of type I errors. Add the false positives to the 80 true positives and you have 125 positive results, fully a third of which are specious. If you dropped the statistical power from 0.8 to 0.4, which would seem realistic for many fields, you would still have 45 false positives but only 40 true positives. More than half your positive results would be wrong.



Click here to watch an animation of this diagram
(http://www.economist.com/blogs/graphicdetail/2013/10/daily-chart-2)

The negative results are much more trustworthy; for the case where the power is 0.8 there are 875 negative results of which only 20 are false, giving an accuracy of over 97%. But researchers and the journals in which they publish are not very interested in negative results. They prefer to accentuate the positive, and thus the error-prone. Negative results account for just 10-30% of published scientific literature, depending on the discipline. This bias may be growing. A study of 4,600 papers from across the sciences conducted by Daniele Fanelli of the University of Edinburgh found that the proportion of negative results dropped from 30% to 14% between 1990 and 2007. Lesley Yellowlees, president of Britain's Royal Society of Chemistry, has published more than 100 papers. She remembers only one that reported a negative result.

Statisticians have ways to deal with such problems. But most scientists are not statisticians. Victoria

Stodden, a statistician at Columbia, speaks for many in her trade when she says that scientists' grasp of statistics has not kept pace with the development of complex mathematical techniques for crunching data. Some scientists use inappropriate techniques because those are the ones they feel comfortable with; others latch on to new ones without understanding their subtleties. Some just rely on the methods built into their software, even if they don't understand them.

Not even wrong

This fits with another line of evidence suggesting that a lot of scientific research is poorly thought through, or executed, or both. The peer-reviewers at a journal like *Nature* provide editors with opinions on a paper's novelty and significance as well as its shortcomings. But some new journals—*PLoS One*, published by the not-for-profit Public Library of Science, was the pioneer—make a point of being less picky. These "minimal-threshold" journals, which are online-only, seek to publish as much science as possible, rather than to pick out the best. They thus ask their peer reviewers only if a paper is methodologically sound. Remarkably, almost half the submissions to *PLoS One* are rejected for failing to clear that seemingly low bar.

The pitfalls Dr Stodden points to get deeper as research increasingly involves sifting through untold quantities of data. Take subatomic physics, where data are churned out by the petabyte. It uses notoriously exacting methodological standards, setting an acceptable false-positive rate of one in 3.5m (known as the five-sigma standard). But maximising a single figure of merit, such as statistical significance, is never enough: witness the "pentaquark" saga. Quarks are normally seen only two or three at a time, but in the mid-2000s various labs found evidence of bizarre five-quark composites. The analyses met the five-sigma test. But the data were not "blinded" properly; the analysts knew a lot about where the numbers were coming from. When an experiment is not blinded, the chances that the experimenters will see what they "should" see rise. This is why people analysing clinical-trials data should be blinded to whether data come from the "study group" or the control group. When looked for with proper blinding, the previously ubiquitous pentaquarks disappeared.

Other data-heavy disciplines face similar challenges. Models which can be "tuned" in many different ways give researchers more scope to perceive a pattern where none exists. According to some estimates, three-quarters of published scientific papers in the field of machine learning are bunk because of this "overfitting", says Sandy Pentland, a computer scientist at the Massachusetts Institute of Technology.

Similar problems undid a 2010 study published in *Science*, a prestigious American journal (and reported

in this newspaper). The paper seemed to uncover genetic variants strongly associated with longevity. Other geneticists immediately noticed that the samples taken from centenarians on which the results rested had been treated in different ways from those from a younger control group. The paper was retracted a year later, after its authors admitted to "technical errors" and "an inadequate quality-control protocol".

The number of retractions has grown tenfold over the past decade. But they still make up no more than 0.2% of the 1.4m papers published annually in scholarly journals. Papers with fundamental flaws often live on. Some may develop a bad reputation among those in the know, who will warn colleagues. But to outsiders they will appear part of the scientific canon.

Blame the ref

The idea that there are a lot of uncorrected flaws in published studies may seem hard to square with the fact that almost all of them will have been through peer-review. This sort of scrutiny by disinterested experts—acting out of a sense of professional obligation, rather than for pay—is often said to make the scientific literature particularly reliable. In practice it is poor at detecting many types of error.

John Bohannon, a biologist at Harvard, recently submitted a pseudonymous paper on the effects of a chemical derived from lichen on cancer cells to 304 journals describing themselves as using peer review. An unusual move; but it was an unusual paper, concocted wholesale and stuffed with clangers in study design, analysis and interpretation of results. Receiving this dog's dinner from a fictitious researcher at a made up university, 157 of the journals accepted it for publication.

Dr Bohannon's sting was directed at the lower tier of academic journals. But in a classic 1998 study Fiona Godlee, editor of the prestigious *British Medical Journal*, sent an article containing eight deliberate mistakes in study design, analysis and interpretation to more than 200 of the *BMJ*'s regular reviewers. Not one picked out all the mistakes. On average, they reported fewer than two; some did not spot any.

Another experiment at the *BMJ* showed that reviewers did no better when more clearly instructed on the problems they might encounter. They also seem to get worse with experience. Charles McCulloch and Michael Callaham, of the University of California, San Francisco, looked at how 1,500 referees were rated by editors at leading journals over a 14-year period and found that 92% showed a slow but steady drop in their scores.

As well as not spotting things they ought to spot, there is a lot that peer reviewers do not even try to check. They do not typically re-analyse the data presented from scratch, contenting themselves with a sense that the authors' analysis is properly conceived. And they cannot be expected to spot deliberate falsifications if they are carried out with a modicum of subtlety.

Fraud is very likely second to incompetence in generating erroneous results, though it is hard to tell for certain. Dr Fanelli has looked at 21 different surveys of academics (mostly in the biomedical sciences but also in civil engineering, chemistry and economics) carried out between 1987 and 2008. Only 2% of respondents admitted falsifying or fabricating data, but 28% of respondents claimed to know of

colleagues who engaged in questionable research practices.

Peer review's multiple failings would matter less if science's self-correction mechanism—replication—was in working order. Sometimes replications make a difference and even hit the headlines—as in the case of Thomas Herndon, a graduate student at the University of Massachusetts. He tried to replicate results on growth and austerity by two economists, Carmen Reinhart and Kenneth Rogoff, and found that their paper contained various errors, including one in the use of a spreadsheet.

Harder to clone than you would wish

Such headlines are rare, though, because replication is hard and thankless. Journals, thirsty for novelty, show little interest in it; though minimum-threshold journals could change this, they have yet to do so in a big way. Most academic researchers would rather spend time on work that is more likely to enhance their careers. This is especially true of junior researchers, who are aware that overzealous replication can be seen as an implicit challenge to authority. Often, only people with an axe to grind pursue replications with vigour—a state of affairs which makes people wary of having their work replicated.

There are ways, too, to make replication difficult. Reproducing research done by others often requires access to their original methods and data. A study published last month in *PeerJ* by Melissa Haendel, of the Oregon Health and Science University, and colleagues found that more than half of 238 biomedical papers published in 84 journals failed to identify all the resources (such as chemical reagents) necessary to reproduce the results. On data, Christine Laine, the editor of the *Annals of Internal Medicine*, told the peer-review congress in Chicago that five years ago about 60% of researchers said they would share their raw data if asked; now just 45% do. Journals' growing insistence that at least some raw data be made available seems to count for little: a recent review by Dr Ioannidis which showed that only 143 of 351 randomly selected papers published in the world's 50 leading journals and covered by some data-sharing policy actually complied.

And then there are the data behind unpublished research. A study in the *BMJ* last year found that fewer than half the clinical trials financed by the NIH resulted in publication in a scholarly journal within 30 months of completion; a third remained unpublished after 51 months. Only 22% of trials released their summary results within one year of completion, even though the NIH requires that they should.

Clinical trials are very costly to rerun. Other people looking at the same problems thus need to be able to access their data. And that means all the data. Focusing on a subset of the data can, wittingly or unwittingly, provide researchers with the answer they want. Ben Goldacre, a British doctor and writer, has been leading a campaign to bring pharmaceutical firms to book for failing to make available all the data from their trials. It may be working. In February GlaxoSmithKline, a British drugmaker, became the first big pharma company to promise to publish all its trial data.

Software can also be a problem for would-be replicators. Some code used to analyse data or run models may be the result of years of work and thus precious intellectual property that gives its possessors an edge in future research. Although most scientists agree in principle that data should be openly available, there is genuine disagreement on software. Journals which insist on data-sharing tend not to do the same for programs.

Harry Collins, a sociologist of science at Cardiff University, makes a more subtle point that cuts to the heart of what a replication can be. Even when the part of the paper devoted to describing the methods used is up to snuff (and often it is not), performing an experiment always entails what sociologists call "tacit knowledge"—craft skills and extemporisations that their possessors take for granted but can pass on only through example. Thus if a replication fails, it could be because the repeaters didn't quite get these *je-ne-sais-quoi* bits of the protocol right.

Taken to extremes, this leads to what Dr Collins calls "the experimenter's regress"—you can say an experiment has truly been replicated only if the replication gets the same result as the original, a conclusion which makes replication pointless. Avoiding this, and agreeing that a replication counts as "the same procedure" even when it gets a different result, requires recognising the role of tacit knowledge and judgment in experiments. Scientists are not comfortable discussing such things at the best of times; in adversarial contexts it gets yet more vexed.

Some organisations are trying to encourage more replication. *PLoS ONE* and Science Exchange, a matchmaking service for researchers and labs, have launched a programme called the Reproducibility Initiative through which life scientists can pay to have their work validated by an independent lab. On October 16th the initiative announced it had been given $1.3m by the Laura and John Arnold Foundation, a charity, to look at 50 of the highest-impact cancer findings published between 2010 and 2012. Blog Syn, a website run by graduate students, is dedicated to reproducing chemical reactions reported in papers. The first reaction they tried to repeat worked—but only at a much lower yield than was suggested in the original research.

Making the paymasters care

Conscious that it and other journals "fail to exert sufficient scrutiny over the results that they publish" in the life sciences, *Nature* and its sister publications introduced an 18-point checklist for authors this May. The aim is to ensure that all technical and statistical information that is crucial to an experiment's reproducibility or that might introduce bias is published. The methods sections of papers are being expanded online to cope with the extra detail; and whereas previously only some classes of data had to be deposited online, now all must be.

Things appear to be moving fastest in psychology. In March Dr Nosek unveiled the Centre for Open Science, a new independent laboratory, endowed with $5.3m from the Arnold Foundation, which aims to make replication respectable. Thanks to Alan Kraut, the director of the Association for Psychological Science, *Perspectives on Psychological Science*, one of the association's flagship publications, will soon have a section devoted to replications. It might be a venue for papers from a project, spearheaded by Dr

Nosek, to replicate 100 studies across the whole of psychology that were published in the first three months of 2008 in three leading psychology journals.

People who pay for science, though, do not seem seized by a desire for improvement in this area. Helga Nowotny, president of the European Research Council, says proposals for replication studies "in all likelihood would be turned down" because of the agency's focus on pioneering work. James Ulvestad, who heads the division of astronomical sciences at America's National Science Foundation, says the independent "merit panels" that make grant decisions "tend not to put research that seeks to reproduce previous results at or near the top of their priority lists". Douglas Kell of Research Councils UK, which oversees Britain's publicly funded research argues that current procedures do at least tackle the problem of bias towards positive results: "If you do the experiment and find nothing, the grant will nonetheless be judged more highly if you publish."

In testimony before Congress on March 5th Bruce Alberts, then the editor of *Science*, outlined what needs to be done to bolster the credibility of the scientific enterprise. Journals must do more to enforce standards. Checklists such as the one introduced by *Nature* should be adopted widely, to help guard against the most common research errors. Budding scientists must be taught technical skills, including statistics, and must be imbued with scepticism towards their own results and those of others. Researchers ought to be judged on the basis of the quality, not the quantity, of their work. Funding agencies should encourage replications and lower the barriers to reporting serious efforts which failed to reproduce a published result. Information about such failures ought to be attached to the original publications.

And scientists themselves, Dr Alberts insisted, "need to develop a value system where simply moving on from one's mistakes without publicly acknowledging them severely damages, rather than protects, a scientific reputation." This will not be easy. But if science is to stay on its tracks, and be worthy of the trust so widely invested in it, it may be necessary.

From the print edition: Briefing