Contents

CHAPTER 1 *'The Same Again, But Different'*: Psychology's Replication Crisis

Hurricane in a teacup

Stop the presses! We bring you some breaking news:

Scientists Say Female Hurricanes are DEADLIER than Male Hurricanes!

Could there be a more twenty-first century headline? Its synergy of human-interest soundbites – gender politics, life and death, even the *weather* – might just represent the epitome of clickbait.

And the news conveyed is so seductively counterintuitive. Or is it intuitive? I suppose it depends on your prejudices. So let us contemplate its main point: according to research, if a hurricane has a woman's name, it becomes *more deadly*. It literally *kills more people*. Hurricane Eve is a greater threat than Hurricane Steve.

But it's a *hurricane*, you might say. How could this be true? How could the name human beings choose for it cause a weather system to wreak a different calibre of havoc? The idea seems, well, implausible.

Nonetheless, the smoke of this news, which made global headlines in 2014, came from the fire of science. Professional researchers reported the claim to be true. More specifically, *psychology* researchers reported it. And psychology, lest we forget, is a science. Psychologists use scientific methods. Psychologists are committed to a search for scientific truth. They wouldn't seek to publish such a finding unless it was reliable, would they?

Well – spoiler alert – as we will go on to see throughout this book, when it comes to psychology, it is never quite safe to assume anything.

The hurricanes-are-deadlier-than-himmicanes study originally appeared in the prestigious American academic journal *Proceedings of the National Academy of Sciences* (Jung et al., 2014). Its authors presented data

showing that more people are killed by female hurricanes than by male ones. A lot more. According to them:

> changing a severe hurricane's name from Charley … to Eloise … could nearly triple its death toll. (p. 8783)

To be fair, they did not claim that the wind blows differently when it has a girly name. They argued that *humans* react differently depending on the gender they perceive a hurricane to be. Humans take female hurricanes less seriously than male ones. It isn't a gender effect so much as it is a gender *bias*. So addled are people by sexist prejudice, they even stereotype the weather. In the words of one CNN columnist (Cupp, 2014):

> *Girl hurricanes get no respect*.

The news spread widely through traditional and social media. And why not? The finding was not just intriguing, it appeared in a highly reputable journal, maybe even one of the most reputable journals in all of science.

But there was, of course, a problem. The finding – while earnestly presented, widely reported, and credulously considered – was not in fact reliable. The claim being made was limited in one crucial respect. It was simply not true.

It did not take long for critics to highlight the many restrictive definitions the authors had used in their analyses (Smith, 2016). They had focused on hurricanes, yes, but this meant they had opted to ignore tropical storms, which are similar and just as deadly. They considered Atlantic hurricanes but not Pacific ones, without an obvious reason for geo-restriction. They confined themselves to hurricanes that made landfall, discounting those that remained offshore, even though offshore hurricanes are regularly just as lethal. (In 2009 the offshore Hurricane Bill became so famous that ten thousand sightseers gathered along the coast of Maine in order to look at it, despite coastguard warnings telling them to stay away. A single wave washed twenty of them into the ocean. Clearly, people didn't give Hurricane Bill much respect, even though it had a male name.)

Most strikingly of all, the authors only considered fatalities within the United States, even though many of the hurricanes they studied affected

neighbouring nations too. In 1980 Hurricane Allen killed 269 people in several Caribbean countries, but only two people in the USA. The authors duly recorded Hurricane Allen as one of the *safest* in their dataset. After all, it killed only two Americans.

The authors had made a series of peculiar choices. They could have included Hurricane Bill, but they opted not to. They could have considered all the fatalities resulting from Hurricane Allen, but they didn't. They could have looked at tropical storms, and events in the Pacific, but they decided to exclude them. Instead they shaved and slimmed their dataset in a way that produced a profound effect on their finding. Their choices made female hurricanes seem more lethal and male ones appear safer.

Statisticians refer to this type of thing as the 'torturing' of a dataset. Such mistreatment of data is as nasty as it sounds. Moreover, as is generally well known nowadays, there are two significant problems with torture. Firstly, it is unethical: it is not morally correct to do it. And secondly, it is ineffective: what gets blurted out under duress might sound good, but, more often than not, it is just not true.

Many researchers have attempted to reproduce the finding that female hurricanes are deadlier than male ones. None have been able to do so. In fact, every single subsequent study has shown that male hurricanes are just as deadly as their female equivalents (Smith, 2016). It turns out that people don't attribute gender stereotypes to the weather after all. They treat girl hurricanes with just as much respect as any other.

Despite making global headlines and remaining on the record as a result of peer-reviewed published research, the finding that female hurricanes are deadlier than male ones is demonstrably specious. Despite many attempts, the result cannot be replicated.

'Fundamentally squishy psychobabble'

It never looks good for psychology when its research findings turn out to be false. It stains the field's image. This is unfortunate, because psychologists are normally quite image-conscious. Professional bodies devote enormous energy to outreach and media work. They develop detailed

strategies to optimize psychology's 'footprint' and 'impact'. They produce packed portfolios of press releases, breathlessly pitching even very vague research summaries as potential news stories. University psychology departments often operate like press offices, feeding stories directly to news outlets, sometimes on a weekly basis. The total collective effort expended in enhancing the field's media profile is little short of monumental.

Presumably, these endeavours are intended to avoid headlines like the following:

> *Study Reveals That a Lot of Psychology Research Really Is Just 'Psycho-babble'.*

Psychobabble is certainly not a term that psychologists would choose to describe their work. But it was the word chosen for them by the UK's *Independent* newspaper in August 2015, when covering a just-published critique of the quality of psychology research (Connor, 2015). The *Independent* succinctly summarized the critique's main finding:

> Psychology has long been the butt of jokes about its deep insight into the human mind… and now a study has revealed that much of its published research really is psycho-babble.

For 'psychobabble' to be mentioned once in a national newspaper is bad enough. For the term to be used repeatedly is certainly not in any psychologist's media strategy.

This psychobabble conclusion was widely reported. The *Guardian* described the finding as 'a bleak verdict on the validity of psychology experiment results' (Sample, 2015). The *New York Times* felt it 'confirmed the worst fears of scientists' regarding the state of psychology (Carey, 2015). According to the *Washington Post*, it confirmed that much of what is published in psychology journals is 'fundamentally squishy' (Achenbach, 2015). Only a modicum of French-language proficiency is required to interpret the verdict of *Le Monde*: 'La psychologie est-elle en crise?' (Larousserie, 2016). By not mentioning psychology in its headline the *Huffington Post* (Reid, 2016) managed to hide the field's blushes (but only just):

> *Scientific Study Proves Scientific Studies Can't Prove Anything.*

Psychologists had been whispering about such a crisis for decades. But now the idea that psychology research is not replicable was about to go viral.

The excitement was stimulated by work produced by the Open Science Collaboration, an international group of over two hundred researchers led by University of Virginia psychology professor Brian Nosek. The group published an extensive and damning report in the elite academic journal *Science* (Open Science Collaboration, 2015). In their programme, they attempted to re-conduct one hundred published psychology experiments. Over 60 per cent of the attempted replications 'failed'.

In other words, the results of the re-conducted studies bore little or no similarity to those of the originals they were based on. In psychology, as in any other science, this is a real problem. When the same study done twice produces two different outcomes, then one thing is for sure: both cannot be right. When 60 per cent of studies are like this, the accuracy of the entire field is called into question. As a column in *Nature* (Baker, 2015), the world's leading scientific journal, bluntly put the point:

> Don't trust everything you read in the psychology literature. In fact, two thirds of it should probably be distrusted.

The problem of course is this: how do you know that the study you are interested in – or indeed any particular psychology research study that you happen to encounter – is one of the trustworthy few and not one of the untrustable majority?

It is little wonder that journalists were interested. Psychology research affects our lives in so many ways. It is consulted in the design of everything from road safety campaigns, to anti-litter initiatives, to health promotion programmes, to educational curricula, to psychometric tests, to household products, to advertising. Politicians cite psychology research when debating policies about childcare, or poverty, or social exclusion, or environmental protection. Even disputes about major constitutional issues, such as whether to extend access to abortion or to lower the voting age, will be influenced by what psychology research has revealed about these issues.

Psychology relates to so much of daily life that its research has universal resonance. If it is actually true that 60 per cent of psychological

science is not replicable – if psychology research is indeed 'fundamentally squishy psychobabble' – then of course the world's media will report this. They have a moral obligation to do so.

According to a famous quotation often attributed to Albert Einstein, insanity can be defined as doing the same thing over and over again and expecting different results. In psychology, such a scenario is the standard way of things. When psychologists conduct the same study over and over again, it would be reasonable to expect *different* results each time, given the historical pattern to date. In Einsteinian terms, it is the confident expectation that a replication might *succeed* that should rightly be considered madness.

The recurrence of non-replication

It would be wrong to think that psychology's replication crisis is new news. In fact, it is very much old news. The Open Science Collaboration itself was the culmination of decades of concern, the outpouring of extreme frustration among an increasingly disaffected subgroup of the psychology research community.

Psychology is often called the 'science of behaviour'. These days, we take behaviour to include people's actions, thoughts, and feelings. Unlike other sciences, psychology is not just a method of inquiry, it is also a field of human activity that itself can be inquired into. Not only can psychologists examine how people in general act, think, and feel, they can also examine how they themselves act, think, and feel. Psychology is itself a behaviour. Therefore, psychology can – and possibly should – study itself.

In historical terms, the use of logic and evidence to uncover knowledge is a relatively modern thing. Before science, and for many centuries, people had little choice but to draw knowledge from authority figures, superstition, or popular consensus. Any claim, no matter how weak, stood a good chance of being believed. People thought that diseases were transmitted through foul-smelling vapours, that all living beings were animated by a spirit-like force, and that California was an island. Science changed all this. It elevated the status of logic and evidence in public discourse. It attached value to validated accuracy. Intellectually, more than technologically, science altered the course of humanity.

Throughout the world, a scientific culture developed during the sixteenth, seventeenth, and eighteenth centuries. Central to science was the idea of objective verifiability. If evidence was to be valued, it had to be verified. If independent observers could not agree that the evidence stacked up, then even once-promising empirical claims would have to be discarded. Verification helped bolster the knowledge base. Falsification – where researchers deliberately sought out evidence that might *disprove* a given claim – bolstered the knowledge base even more. The core tenet was that science not only established the facts, it sought to check them and then to re-check them, and then to re-check them again.

By most accounts, psychology became a mainstream science during the middle of the nineteenth century. This means that psychological research has been conducted for over one hundred and fifty years. Literally millions of papers have been published. Given science's emphasis on objective verifiability, we might reasonably assume that a huge volume of this work has involved the checking and re-checking of previous findings. However, this is simply not the case. According to one analysis, only 1 per cent of papers published in the top one hundred psychology journals *during the last one hundred years* have been replications (Makel, Plucker, and Hegarty, 2012). Replications in psychology are not just uncommon. They are almost unheard of.

Nonetheless, given the overall volume of research during this period, 1 per cent still amounts to thousands of studies. Therefore, it should be possible to evaluate the effectiveness of these replications. Computations of psychology's replication rate have thrown up two main conclusions.

Firstly, quite often published replications confirm the results of the original studies on which they are based. But secondly, a worrying pattern emerges when it comes to study authorship: replications are likely to corroborate previous findings when they are performed by *the same researchers who conducted the original studies*. They are much less likely to do so when conducted by independent investigators (Makel, Plucker, and Hegarty, 2012).

In other words, when psychologists check their own homework, they tend to do very well. But when someone else checks it for them, they tend to flunk.

All this raises two important conclusions regarding replication in psychology. The first is extremely disappointing, the second extremely worrying. The disappointing conclusion is as follows. Despite the

importance of objective verification and replicability in science, psychology clearly lacks a replication culture. Psychology can hardly convince audiences that its findings are replicable when custom and practice is to avoid even *attempting* to replicate research.

The worrying conclusion is that this reluctance to replicate may be hiding a much bigger problem. The link between study authorship and successful replicability – where replications 'succeed' when conducted by friends but 'fail' when conducted by strangers – raises questions about scientific objectivity. It closely resembles what would likely happen if researchers deliberately biased their studies to produce outcomes they particularly wanted.

For many years, this was psychology's dirty little secret. This is no longer the case. That is, it is no longer a secret.

The prehistory of psychology's replication crisis

The idea that psychology might have a problem with replication is as old as psychology itself. As soon as psychology emerged during the late nineteenth century, there was concern that what was taught might not be reliable.

The first psychology departments in Europe and North America were effectively repurposed religious philosophy departments: contemplation of the soul had gradually morphed into research on the mind (Fuchs, 2000). The Victorian fascination with psychics and séances had a strong influence, even on the most rarefied academic institutions. In the late 1800s Harvard University's psychology department listed a Ouija board in its inventory of laboratory equipment (Coon, 1992). The tenuousness of such subject matter proved disconcerting to many.

Leading psychologists feared ridicule were psychology to become sucked into the world of esoteric ephemera. Many mediums whose powers were praised with earnest sincerity by eminent scholars turned out to be charlatans (Coon, 1992). So extensive were the problems, psychologists quickly recognized that unguarded observations were too naïve to be relied upon for research purposes. Formal scepticism was required. As a result, several now-standard techniques for strengthening research – including blinding and counterbalancing – were developed during this period (O'Keeffe and Wiseman, 2005).

The late nineteenth century also saw the emergence of Freudianism, and with it the psychoanalytic school of psychology. While psychoanalysis proved extremely popular, it also attracted suspicion from the scientifically minded. A core tenet of psychoanalysis was the claim that the unconscious – an entity that by its very definition was unobservable – truly existed. According to critics, such a theory was the opposite of science. It completely disregarded the principle of objective verifiability. It generated a quagmire of open-ended prognostications. Concerns about whether psychoanalysis could ever produce replicable findings quickly became widespread.

By the twentieth century psychologists had become very exercised about rigour. There emerged a prevailing view that they should only concern themselves with observable – and thus replicable – phenomena. Figures such as Fechner and Wundt developed the field of psychophysics, which used standardized methods to examine the limits of human senses under laboratory conditions. By studying the dimmest lights that people could see or the smallest differences in volume that they could hear, the psychophysicists established that sense discrimination was mathematically predictable. Psychophysics was premised on the notion of replicability: data were informative only insofar as they were consistent (more or less) across all humans who were studied.

Knowledge arising from psychophysics could be helpful in certain contexts, but its applicability was limited. Its focus on exactitude served to restrict the range of subjects that could be studied. Psychology was much more than psychophysics, and it required much more powerful approaches.

Many psychologists formed the opinion that a better way to study human experience was to focus on externally observable behaviours. This led to the emergence of behaviourism, a theoretical view that dominated psychology for much of the twentieth century. Its rapid spread reflected pervasive unease about the replicability of psychological observations. Like the psychophysicists, the behaviourists strove to focus on quantifiable data and on whether what they saw in one human (or laboratory rat) could be generalized to others.

Psychophysics and behaviourism loom large in almost every telling of the history of psychology. Both emerged in response to anxieties about non-replicability. These fields spearheaded an emphasis on scientific rigour that now exists across psychological research, but they also highlight the fact that non-replicability is by no means an exclusively modern concern.

The problem of Rampant Methodological Flexibility

This exponential growth in scientific psychology brings its own challenges. At any given moment, thousands of investigators conduct tens of thousands of studies, the results of only some of which will eventually see the light of day in print. There is no one standard way to conduct psychology research. Much of what is planned and executed is devised by the researcher concerned and so is subject to whatever choices or preferences the researcher has at the time. There has emerged in psychology a culture of Rampant Methodological Flexibility.

Rampant Methodological Flexibility makes it extremely difficult to compare different studies. While two researchers in separate laboratories might conduct research on the same topic, their methods will inevitably deviate. Whether one study truly replicates the other is unclear. If both studies produce similar results, the investigators can claim to be corroborating each other. But if the studies produce different results, the investigators can comfort themselves by attributing the divergence in findings to differences in research methods.

Researchers allow self-serving bias to addle their scientific brains. When faced with ambiguous results, it is part of human nature to think of your own study as being well designed, coherent, and worthwhile. Even the most conscientious researcher will be inclined to interpret data in a way that bolsters this self-flattering premise. Whether intentionally or unconsciously, researchers end up engaging in a practice now often called *HARKing* – 'hypothesizing-after-the-results-are-known' (Kerr, 1998). They employ logical fudges to make arbitrary sense of ambiguous findings.

Rampant Methodological Flexibility also allows researchers to make on-the-hoof decisions about how studies are conducted. They are free to nudge their results in desired directions. For example, if they check the data from their first twenty participants and see what they like, they can continue the study and produce the results they want. However, if they don't see what they like, they have options. They can discontinue the study and design a new one. Or they can modify the existing study and continue the research, having tweaked the procedure in a desired manner. Or they can decide, retrospectively, to rewrite their study 'prediction' so that their data no longer contradict their hypothesis. In effect, if they don't see what they like, they can choose to like what they see.

Ultimately, by the time the research is published in a scientific journal, all such discretionary methodological choices will be part of dim and distant history. Most will unlikely ever be reported. Readers of scientific journals are faced with thousands of studies that turned out the way they did because of self-serving researcher influence but that could have turned out differently if *circumstances* had been different. Readers only see studies that make it into press. They don't see the studies that were abandoned or those that would have taken place had the researchers made different design choices.

Astute psychologists noticed this problem many years ago. In 1967 Paul Meehl, a professor at the University of Minnesota, warned of 'eager-beaver researcher[s], undismayed by logic-of-science considerations', who wend their way through strings of ad hoc methodological adjustments, publishing study after study but producing findings that simply cannot be relied upon (Meehl, 1967). Meehl saw this problem as extending far beyond self-serving bias among investigators. He described a structural problem that affected psychology as a whole. Psychology did not just *tolerate* methodological flexibility; it actively rewarded and encouraged it. His observations were published in the niche academic journal *Philosophy of Science* half a century ago. Nearly fifty years later the Open Science Collaboration – and the world's media – were reporting that not that much had changed.

The modern replication crisis

At the turn of the twenty-first century murmurs about psychology's structural replication problem became a burgeoning hubbub of intrigue and ignominy. In some ways psychology had changed. The field had grown to overlap with sister disciplines such as sociology and biology. Its subject-area remit widened to incorporate a range of social and cultural issues, as well as some hard-core neuroscience. But in other respects psychology had stayed the same. The new topics and approaches did little to alter the problems caused by Rampant Methodological Flexibility.

Indeed, in many ways they had made them worse. For example, brain imaging researchers discovered they were getting their sums wrong. Their conventional statistical approaches grossly inflated results, leading

to entirely spurious findings in many cases (Vul et al., 2009). The problems seen in imaging studies also hampered research on event-related potentials, another technology used to measure brain activity (Luck and Gaspelin, 2017). Around half of all papers in top-tier neuroscience journals were found to report statistical interactions ineptly, generating waves of false conclusions about their meaning (Nieuwenhuis, Forstmann, and Wagenmakers, 2011).

Contributing to these problems was the way in which neuroscientists apply high-powered statistical procedures to low-powered tiny studies. It is very much a case of using sledgehammers to crack nuts. Simply put, most brain imaging studies use samples too small to support confident statistical conclusions. According to one assessment, brain imaging studies are so minuscule they can be expected to find no more than eight out of every one hundred real effects that exist (Button et al., 2013). This means that, for every hundred effects described in the literature, as many as ninety-two are liable to be flaky. They are likely to belong to the category of 'false positives': outcomes that *look* like findings but that aren't really findings at all, just freakish statistical patterns produced by chance.

In the years leading up to the Open Science Collaboration, several other investigations were casting different kinds of doubt on psychology. Questionable research practices were shown to be common. In one survey the authors of over 140 studies appearing in major psychology journals were asked to share their datasets (Wicherts, Bakker, and Molenaar, 2011). Even though all had previously signed agreements to do so (a requirement for research to be published), the majority now refused. Researchers who had reported the most tenuous findings were most likely to decline, suggesting an awareness of weaknesses in their work.

Statisticians used modern analytic techniques to gather direct evidence of the problem of Rampant Methodological Flexibility. Their approach was to scrutinize large swathes of published psychology research and to compare its findings to statistical models of what would have been produced had the science been truly unbiased. This work revealed an overall statistical pattern of results that was so improbable, it implied a large proportion of published findings in psychology are either cherry-picked, manipulated, erroneous – or simply made up (Leggett et al., 2013).

The prospect of fraud should not be dismissed lightly. Survey results suggest that around one in seven psychologists are directly aware of

colleagues who have falsified data (Fanelli, 2009), and a number of particularly audacious fakery cases, some in very high-profile research groups, made international news headlines in the early 2000s (Estes, 2012).

But not all factors leading to non-replicability relate to researcher misbehaviour. Some reflect adverse attitudes among the intended gatekeepers of scientific standards. The fact that replications appear so rarely in psychology journals is only partly due to the lack of studies. To a greater extent, it reflects the reluctance of editors to consider such papers for publication.

In late 2015 the Open Science Collaboration published the long-awaited findings of its 'Reproducibility Project'. Over a three-year period lead author Brian Nosek had persuaded 269 colleagues from around the world to join him in attempting to replicate one hundred studies from major psychology journals. To focus on a fair selection of research, Nosek randomly drew studies from a single year (2008) and from three top-tier journals, namely: *Psychological Science*, the *Journal of Experimental Psychology: Learning, Memory, and Cognition*, and the *Journal of Personality and Social Psychology*. He then categorized the selected studies as either social psychology (i.e., those from the *Journal of Personality and Social Psychology* and half of those from *Psychological Science*) or cognitive psychology (i.e., those from the *Journal of Experimental Psychology* and the remaining ones from *Psychological Science*).

Detailed protocols were established for constructing and conducting the replications. Investigators were required to contact the original authors for study materials and to invite them to review the replication methods. To avoid the risk of HARKing, investigators also had to publicize their methods before commencing their replication. After all the replications were completed, their results were compared with those of the originals. Nosek and his colleagues considered two types of comparison. First, they looked at statistical significance – in other words, whether the replication yielded evidence in support of the original research prediction. Then they looked at the size of the statistical effects – in other words, whether the replicated result was similar to the original not just in trend but also in intensity or impact.

As we know, the picture was wholly underwhelming. While 97 per cent of the original studies had produced statistically significant findings, only 36 per cent of the replications did so. Replicability in social

psychology seemed particularly poor. Only 23 per cent of studies from the *Journal of Personality and Social Psychology* were successfully replicated, along with just 29 per cent of the social psychology studies from *Psychological Science*. While cognitive psychology proved more replicable, its record was still far from stellar. Only around half of these replications were successful (48 per cent of those from the *Journal of Experimental Psychology* and 53 per cent of cognitive studies from *Psychological Science*). Nosek and his colleagues then performed computations to calculate a score, on a scale of 0 to 1, that represented the effect size of the various studies. The median effect size of the original studies was 0.403; that of the replications was just 0.197. This showed that even where studies seemed replicable, replications produced much weaker findings than the original research.

The picture is clear. Psychology has a problem with replication. This problem is not restricted to social psychology, which is often said to be difficult to study scientifically. It also affects cognitive psychology and neuroscience, both said to be very scientific. The problem is not confined to research in obscure journals with dubious editorial standards; it emerges throughout top-tier journals also. And it is not just the result of self-serving bias among researchers; it is exacerbated by editorial practices that encourage Rampant Methodological Flexibility and discourage the conducting of replications.

The problem takes two main forms. Firstly, psychology flagrantly lacks one of the key hallmarks of science – a *culture* of replication. And secondly, large numbers of psychology studies fail the basic test of science – under the challenge of replication, they fail.

Non-replicable findings are everywhere

Psychologists like the term *confirmation bias*. It describes how people always try to prove themselves right. When judging an unproven proposition that they like, most people are selective when it comes to the evidence they look for. They are also selective in interpreting the evidence they find.

People are rarely truly objective. They over-focus on information that supports their prior beliefs, and they generously interpret ambiguous

information as if it was clearly in their favour. In short, they engage in logical fraud. They pretend the evidence is more conclusive than it actually is, and they congratulate themselves on their perceptiveness in having been proved 'right'. To be fair, most people do this somewhat unconsciously, and this is what is properly referred to as *confirmation bias*. When it is done consciously, it is good old-fashioned cheating.

Psychologists have conducted thousands of research studies on confirmation bias, so you would think they would know lots about it by now. They should certainly be familiar with the issue. This is because as well as studying confirmation bias, they also succumb to it. When poor quality psychology research is published or cited widely, it is nearly always because of confirmation bias.

Research is considered much more plausible when researchers, reviewers, editors, and ultimately readers believe that the findings 'make sense'. As such, results that support a commonly held view, or a fashionable theory, or an obscure principle held dear by technical specialists are much less likely to be intensively scrutinized. Less attention is paid to the methodologies used to produce them.

It is very common for weak research papers to receive generous evaluations because editors and reviewers become swayed by the expectation that studies should produce 'logical' results. Female hurricanes are more deadly? Well it makes sense; after all, implicit sexism is an important social problem. Psychology journals accumulate many papers because of such confirmation biases, perhaps more than are accepted because of compellingly sound science.

Take, for example, a 2011 study in the journal *Applied Animal Behaviour Science*, which claimed to show that dogs are four times more likely to bite other dogs when they are walked by a man than when they are walked by a woman (Řezáč et al., 2011). The implication is that male dog-walkers, by virtue of being men, somehow transmit their masculine aggression to the dogs they walk. It was widely reported that the study involved the observation of nearly two thousand dogs and their owners: the Discovery Channel referred to it as 'one of the world's largest studies of dogs on walks' (Viegas, 2011). However, the research was weak in several respects.

Firstly, it was an observational study in which, across several weeks, student research assistants lurked in public parks with clipboards and

watched people walking their dogs. Whether the same dogs and walkers were included on multiple occasions was not accounted for. Secondly, the study was cross-sectional, so it is impossible to infer cause-and-effect relationships. Maybe more aggressive dogs are walked by men precisely *because* they are more aggressive. In other words, maybe there is a tendency among dog-owning families to ask their menfolk to walk the more troublesome canines. Thirdly, the data were analysed in a very simplistic way. The authors used bivariate instead of multivariate techniques. This means that they did not seek to statistically control for multiple other potential causes of dog aggression, such as dog size, when analysing gender differences. And finally, the study was far from one of the world's largest studies of dogs on walks. In fact, it might have been one of the smallest. For sure, around two thousand dogs were observed in the research as a whole. However, only twenty-eight were seen biting (this was the authors' working definition of dog aggression). As such, only this subgroup of *twenty-eight* – effectively comprising a mini-study within the overall research project – could be used to investigate the impact of gender-of-walker on dog aggression. By any standard, this is a fatally insufficient sample. It seems that confirmation bias drove the publication of this study and its coverage in the news media. The study's many weaknesses, including the tiny sample size underlying its headline finding, were overlooked. To readers, the idea that male dog-walkers make their dogs more aggressive seemed to 'make sense'. It conformed to a cultural stereotype that depicts men as inherently more violent than women.

In recent years an increasing number of initially attention-grabbing psychology studies have ended up foundering due to this confirmation bias problem. Many relate to claims about personal fate and well-being. In one very prominent example, researchers asserted that adopting a 'power pose' – standing with your hands on your hips and your elbows pointing outwards, like a comic-book superhero – produces several immediate benefits, including enhanced feelings of competence, increased willingness to take risks, lower levels of the stress-hormone cortisol, and higher levels of testosterone (Carney, Cuddy, and Yap, 2010). One of the authors delivered a TED talk based on the research, which has now been viewed over 40 million times. In it she claims that power poses are so impactful, people who use them 'can significantly change the outcomes of their life' in just two minutes (Cuddy, 2012).

In some ways this seems like quite an implausible notion, one that few people would have anticipated in advance, and so immune from confirmation bias per se. However, the idea that people can exert agency over their own destiny is a very strong cultural belief (Yarritu, Matute, and Vadillo, 2014), as is the notion that physical posture is associated with inner psychological character (Gilman, 2014). This, coupled with the Ivy League affiliations of the study authors and the esteem of the publishing journal, created perfect conditions for confirmation bias. In short, readers of the research quickly concluded that, all things considered, the findings 'made sense'. The problem, of course, is that they really didn't. The study was based on a sample of just forty-two participants. When independent investigators conducted replications based on samples of two and three hundred, they found no power pose effects whatsoever on any of the alleged outcome variables (Garrison, Tang, and Schmeichel, 2016; Ranehill et al., 2015). The original power pose findings have been thoroughly debunked, so much so that the lead author published a statement on her personal website repudiating them in forthright terms (Carney, 2016). It turns out it takes more than two minutes of standing like a superhero to improve your life outcomes.

Similar doubts have surfaced concerning other research linking simple behaviours to complex consequences. These include a famous study claiming to show that moving your facial muscles into the shape of a smile – such as when you hold a pencil sideways in your mouth – causes neurological feedback that engenders feelings of happiness (Strack, Martin, and Stepper, 1988). Notwithstanding the extent to which this study is cited in introductory psychology textbooks, it turns out the reported effect cannot be replicated (Wagenmakers et al., 2016).

A comparable controversy surrounds research on the concept of 'ego depletion', an effect in which people's willpower is shown to be a finite resource that declines the more it is used (Baumeister et al., 1998). The original study has been extensively cited, and several researchers have claimed to have demonstrated ego depletion in different experiments. However, critics have long questioned whether it really exists. One analysis of over a hundred ego depletion experiments showed their combined statistical result to be effectively zero (Carter et al., 2015). A subsequent literature review found that ego depletion studies show a bias towards false-positive (i.e., random) effects (Schimmack, 2016). More damningly, a direct replication of the original study, conducted by

independent investigators using a much larger sample, found no evidence for ego depletion at all (Hagger et al., 2016).

In another famous study, researchers claimed that participants walk more slowly after being presented with words like 'bingo' and 'Florida', words that the study's authors argue plant stereotypes of the elderly in participants' minds (Bargh, Chen, and Burrows, 1996). The finding stumbled and fell upon replication (Doyen et al., 2012).

Other researchers have claimed to show that getting people to hold hot cups of coffee makes them more likely to rate others as generous and caring (Williams and Bargh, 2008). The effect implied that making people physically warm also makes them interpersonally warm. A replication based on a much larger sample was sufficient to pour cold water on that conclusion (Lynott et al., 2014).

A related line of research has attempted to link cleanliness with morality. While attracting positive attention for many years, the field is in disarray now that replications have shown the effects to be untrustworthy. One experiment claimed that handwashing makes people more forgiving (in other words, that hygienic purity engenders a sense of moral purity; Schnall, Benton, and Harvey, 2008). Another claimed that threatening people's sense of morality makes them seek out ways to cleanse themselves (in other words, that moral purity engenders a craving for hygienic purity; Zhong and Liljenquist, 2006). Both widely cited experiments have been subjected to robust replication attempts (Earp et al., 2014; Johnson, Cheung, and Donnellan, 2014). In both cases the original findings have been washed away.

Research that cannot be replicated comes in all shapes and sizes. A study suggesting that people pursue professions that remind them of their own name (e.g., women named Laura are more likely to become lawyers, while men named Dennis are more likely to be dentists; Pelham, Mirenberg, and Jones, 2002) is cited in several introductory psychology textbooks. However, according to a replication, the effect is actually spurious (Simonsohn, 2011). A related finding that people are more inclined to marry someone whose name is similar to their own (making Josephine more likely to plight her troth to Joseph; Jones et al., 2004) also disappeared when the relevant work was replicated (Simonsohn, 2011).

An influential study of day-old babies, which was supposed to show that baby boys liked staring at mechanical mobiles whereas baby girls preferred to gaze at human faces (Connellan et al., 2000), saw its finding

vanish in a replication attempt (Leeb and Rejskind, 2004). A replication of another classic baby study (Meltzoff and Moore, 1977) showed that newborns do not, after all, possess the ability to imitate facial expressions of adults (Oostenbroek et al., 2016).

Seeing a picture of a pair of human eyes (e.g., as part of a sign) was originally reported to make people engage in more prosocial behaviour (Bateson, Nettle, and Roberts, 2006), and at least one police force adopted the idea when designing its anti-crime posters. But an independent attempt to replicate the finding in a larger sample found that, actually, no such effect occurs (Carbon and Hesslinger, 2011).

This is just a small selection of non-replicable research in psychology. In the main, these studies are well known: most are regularly cited in textbooks, and many are considered classics. When replications cause the core of the undergraduate curriculum to collapse like a house of cards, it becomes clear that psychology has a serious problem.

It should be recalled that replications make up only 1 per cent of published psychology research. For every high-profile study that attracts a replication, there are hundreds of others, most of them banal, whose replicability remains unknown. The true extent of non-replicability in psychology is a matter of speculation. But we can be sure that it is likely to be huge.

Why is psychology at risk?

So why is psychology's replicability problem so chronic? One way to answer this is to consider the factors that make spurious findings more likely. For example, we know that small studies are more likely to produce false-positive results than large studies. Therefore, if a field is dominated by small studies, its rate of false-positive results will be high. Similarly, we know that fields that study small effects (e.g., subtle group differences that are hard to notice without the aid of fine-grained statistical information) are more likely to produce false positives than fields that study large effects. Therefore, if a field is dominated by the study of small effects, then *its* rate of false positives will be high.

We know that taking a scattergun approach to statistical analysis – testing everything that can be tested in the hope of finding something statistically interesting – also heightens the likelihood of false-positive

results. Therefore, if a field is characterized by such scattergun approaches, then it too can be expected to produce very high rates of false-positive findings.

And fourthly, if research in a given field is conducted with a high degree of methodological flexibility, wide-ranging analytic approaches, and inconsistent definitions of key terms, then *its* rate of false-positive findings can also be expected to be very high. Therefore, a field whose norm is Rampant Methodological Flexibility can be expected to produce a torrent of spurious, unreliable, and non-replicable findings.

The bad news is that psychology possesses not just some but *all* of these characteristics. Indeed, in addition, psychology possesses several other features that undermine research accuracy (such as: a proneness to study fashionable, but not well-understood, subject matter; a proneness to study topics about which researchers may hold personal biases or prejudices; and a culture of hectic competition among researchers to be the first to publish findings on a given topic).

One medical statistician has attempted to quantify the impact of all such factors in order to compute the general likelihood of spurious findings in the biological and health sciences (including psychology). According to his analysis, the odds of false positives are greater than fifty-fifty. Therefore, as he puts it, 'most claimed research findings are false' (Ioannidis, 2005).

Other statisticians have noted that studies published in scientific journals represent only a subset of those that have actually been conducted in the world. Scientists often commence studies that they eventually choose to abandon or abort. This usually happens when they conclude that a study's finding is trivial, nondescript, or otherwise boring. Quite often it is because the finding is not so much a finding as it is a non-finding: the predicted effect didn't materialize, or the anticipated group difference just wasn't there. In other words, the researcher's original conjecture was wrong. Rather than publish such a non-finding, the researcher discards the study, moves on, and invests their effort in designing new research.

The problem is that this serves to filter banal findings from the published literature. Readers who study the contents of journals are left with a skewed impression of the evidence. They will see plenty of statistically significant findings (the majority of which, according to Ioannidis, will be spurious). But they will *not* see those studies on the same topics that

were abandoned because the researcher found nothing of interest in the data. In other words, published research on a given topic is difficult – if not impossible – to interpret unless you also have full information about the *unpublished* research that has been conducted on the same topic. This quandary was noted very long ago by an American psychologist called Robert Rosenthal. In reference to the way in which researchers stored their old unwanted papers at the time, he described it as the 'file-drawer problem' (Rosenthal, 1979).

Nowadays, the extent to which published research represents a biased selection of all research can be inferred statistically. As we will see in Chapter 4, psychology research typically determines the importance of findings based on statistical computations of probability. If the data collected are found to be improbable (allowing for various assumptions), then the finding is considered to be of interest. More specifically, if the statistical test shows this improbability to be less than one in twenty (i.e., if its '*p*-value' is less than 0.05), then the result is declared 'statistically significant'.

Strictly speaking, as psychology research looks at natural phenomena, the probabilities computed for the data in such studies should reflect patterns seen in nature. The numbers should be spread like raindrops on a window pane or leaves on a tree: any vividly recurring pattern or exact right-angles would suggest human interference. However, when statisticians scrutinize the shape of psychology findings, they find many curious contours. The patterns deviate from what should be seen in a fair selection of research.

For example, statisticians have observed an inexplicable spike in *p*-values lying just below the magic 0.05 level (Leggett et al., 2013). In reality, there is no reason for the probability of naturally occurring data to cluster around a point just below 0.05 (or indeed around any other number). The fact that psychology data do this suggests *either* that they have been analysed in idiosyncratic ways in order to contrive this very outcome *or* that the studies themselves are highly selected and do not represent a fair sample of what has been done. Actually, both scenarios could be true.

Perhaps the biggest factor producing so many spurious findings is Rampant Methodological Flexibility. Unlike many other sciences, psychology lacks a standard experimental procedure. There is no single

approved approach to research. Even studies that use comparable methods to examine similar questions will end up distinct from one other. Individual researchers are free to customize their research designs, alter their procedures, tweak their analytic practices, redefine their terms, and even retrospectively rewrite their hypotheses. This level of flexibility helps make psychology an adaptable field that can be tailored to any topic or research circumstance.

However, Rampant Methodological Flexibility also opens the door, extremely widely, to confirmation bias. It allows dishonest researchers to distort studies to suit their own needs, corroborate their own predictions, or otherwise make themselves look good. Honest researchers can also be tempted to do these things, albeit unintentionally. Whether researchers are dishonest or honest makes little difference to replicability. By facilitating confirmation bias, Rampant Methodological Flexibility makes it almost certain that psychology will produce plenty of spurious findings that simply can't be replicated.

The problems are compounded by commonplace scholarly weaknesses. Psychology is full of weak statistical practices, weak research methods, and weak theory. Weak statistical practices include a near-obsessive focus on probability values produced by statistical tests. As we will see in Chapter 4, researchers in psychology regularly fail to apply statistical tests correctly, do not fully understand the nature or purpose of such tests, and habitually misinterpret their results. Further, many psychologists – if not most – make unsound distinctions between findings based on whether the associated *p*-values are below or above the 0.05 threshold. This near-obsession with categorizing *p*-values leads to several erratic approaches to analysis (such as those that produce a bulge in *p*s just below 0.05). Rampant Methodological Flexibility allows researchers to try any number of statistical approaches in an effort to test and retest the same dataset to hunt for a suitable *p*. Whenever *p* falls below 0.05, an expedient researcher can choose there and then to cease the study, desist from any further analysis, and punch the air in celebration. Indeed, the culture of celebrating statistical significance as if it were some kind of victory (leading many a disappointed student to ask their supervisors 'What went wrong?' when their *p* is too high) is itself a serious problem.

The belief that the only good finding is a statistically significant one pressurizes researchers to exploit methodological flexibility in ways

that undermine objectivity. Needing to resort to idiosyncratic statistical analyses in order to shake a significant *p*-value from an otherwise moribund dataset virtually guarantees that the finding so produced will be non-replicable.

Psychology's weak research methods include a decades-long and incorrigibly persistent reliance on small study samples. As mentioned above, sample sizes in brain imaging studies are, on average, so small that only eight out of every hundred real effects can reliably be observed (Button et al., 2013). In psychology as a whole, average sample sizes are large enough to detect only twenty-four effects out of every hundred (Smaldino and McElreath, 2016). Given that the vast majority of research papers report statistically significant findings, this could suggest that between 80 and 90 per cent of them are likely to be false positives (we will consider this matter further in Chapter 5). The fact that psychologists are often sanguine about the sample-size problem may reflect a generally casual attitude to the broader issue of research methods.

While research methods training is a core requirement of psychology degree education, it is generally considered socially acceptable among psychologists (especially among applied practitioners) to be somewhat disregarding of their importance. Some movements within psychology have adopted scepticism toward quantitative methods as their badge of honour. Subfields such as critical psychology and qualitative psychology not only argue that statistical methods are inessential, some voices can be heard claiming they are so ill-suited to the study of human behaviour as to be a blight on the social sciences. Given this culture, perhaps it is no surprise that some psychologists fail to prioritize statistical competence when attempting to conduct research.

Weak theory is also problematic. In its pure form, science seeks to build future knowledge by methodically refining present knowledge. It inches ahead in incremental steps, plodding along, objectively oblivious to the limelight. In contrast, modern psychology is often keen to take large bounding strides forward, in bold and newsworthy ways. It seems that many top-tier psychology journals prioritize the publication of studies that can be deemed 'exciting' or 'quirky', giving high-profile platforms to such notions as power poses, priming, and precognition.

However, the more novel the conceptual content of research, the greater its risk of relying on unproven prior assertions. Rather than standing on

the shoulders of giants, such studies attempt to levitate high in the sky in order to see more and further than other studies can. But the problem with wanting to levitate is that it is untenable, and probably impossible. Some distortion of reality will be required.

But is psychology's replication crisis overstated?

In any crisis, there are the people who ask 'Crisis? What crisis?' Psychology's replication crisis is no exception. Almost as soon as the Open Science Collaboration published the results of its reproducibility project, there were critics who complained that the work itself was an example of spurious research in psychology. In the *Times Higher Education*, social psychologists Wolfgang Stroebe and Miles Hewstone (2015) argued that the Collaboration's results were as untroubling as they were unsurprising. They noted that single studies are always weaker than the combined results of multiple studies, so therefore conducting single replications was not a good way of testing original experiments. Instead, they said, combined analyses of many studies (so-called meta-analyses) should be considered the gold standard. According to Stroebe and Hewstone, many such meta-analyses had shown social psychology findings to be reliable. They dismissed the Open Science Collaboration findings as 'meagre' and 'not very informative'.

Perhaps the highest-profile scepticism directed at the Open Science Collaboration came from Daniel Gilbert, a Harvard University social psychologist. His critique appeared in *Science*, the same journal in which the Open Science Collaboration results had been published (Gilbert et al., 2016). His main complaint was that the replications were not sufficiently similar to the original studies on which they were based. For example, a study on racism originally conducted in the United States was replicated in the Netherlands. Gilbert argued that sociopolitical differences between the two countries rendered the Dutch replication problematic, and so it was unsurprising its results failed to match those of the American original. According to Gilbert, when such methodological differences are factored into analyses, the replication rates observed by the Open Science Collaboration fall well within expectations.

Echoing Stroebe and Hewstone, Gilbert further argued that conducting single replications for each original study was an insufficient way to test for false positives. After all, if the two findings differ, it does not necessarily follow that the original study is the one that is spurious. Maybe the replication result is the outlier. By failing to allow for this possibility, Gilbert suggested, the Open Science Collaboration inadvertently tested the original studies against artificially high benchmarks. Gilbert also argued that the Collaboration failed to account for bias. He suggested that differences in methodologies between original studies and replications might reveal that the investigators who conducted the replications were biased (either explicitly or implicitly) and secretly willed their replications to fail.

The Open Science Collaboration published a response to Gilbert's criticisms (Anderson et al., 2016). They claimed that Gilbert's critique contained several inaccuracies and statistical misunderstandings. They reported that several of the replication methodologies Gilbert criticized were approved by the original studies' authors. They pointed out that Gilbert omitted to discuss a number of their findings that directly contradicted his critique. Finally, they turned Gilbert's point about study differences on its head. They agreed with Gilbert that the replication studies differed from the originals they were replicating. Given the fact that replications are conducted at different times and by different people, they noted that 'there is no such thing as exact replication'. As such, while Gilbert had argued that the Collaboration tested the original studies against artificially high expectations, the Collaboration responded by arguing that Gilbert had critiqued their replications against similarly inflated standards.

It is sometimes noted that problems with replication are not unique to psychology. This is certainly true. In medicine the development of rigorous protocols for randomized controlled trials (RCTs) was itself a response to widespread concerns about non-replicability in medical studies. Anxiety about the reproducibility of drug efficacy trials dates back to the 1950s (Bastian, 2016). As recently as 2012, an attempt to replicate the findings of fifty-four 'landmark' preclinical cancer research studies found that only six could in fact be replicated (Begley and Ellis, 2012). So psychology is not alone. But of course, multiple wrongs do not make a right. Just because other fields have replication crises of their own does not make psychology's any less grave.

Models or muddles?

In reality, concerns about psychology long predate the Open Science Collaboration. Whether or not their reproducibility project has shortcomings, we know that the crisis exists. We *know* that psychology lacks a replication culture: replication accounts for just 1 per cent of published research. We *know* that replications are less likely to succeed when they are conducted independently. We *know* that Rampant Methodological Flexibility opens the door to confirmation bias. We *know* that psychology frequently touches on subject matter about which researchers will hold strong personal opinions, thereby straining their ability to remain objective. We *know* that the file-drawer problem exists and that, therefore, published studies overstate the commonness of effects. We *know* that sample sizes in psychology are statistically too tiny to inspire confidence in research. We *know* that effects in psychology are often very subtle and so risk flooding the literature with spurious false-positive findings.

We *know* that researchers wilfully engage in questionable research practices. We *know* that scattergun approaches to data analysis are the norm rather than the exception. We *know* that researchers can retrospectively rewrite their hypotheses once their results are known. We *know* that several landmark studies have been cited in textbooks for years, even though their findings cannot be reproduced by independent investigators. We *know* that editors and readers overlook methodological limitations when study results are deemed to 'make sense'. We *know*, in other words, that psychology is at high risk of suffering a perennial problem of non-reproducibility and that radical action is needed.

Psychology's problems with replication encompass multiple methodological crises. In the rest of this book, we will examine each in turn. Psychology has significant challenges with the task of measurement. If psychological concepts are not quantified properly, then findings based on such measures will be hard, if not impossible, to replicate. Psychology faces many problems with statistical analyses. Quite how best to apply statistical techniques to the types of data gathered in psychology research is a matter of constant debate. Psychology has a particular challenge with sampling. How exactly do psychologists ensure that the people they study, and the situations in which they study them, yield findings that

can be generalized to humanity as a whole? When the findings are in, psychology then faces severe difficulties with interpretation. The temptation to fill explanatory gaps with subjective judgements creates more confusion than certainty.

Psychology also faces problems with its own coherence. In many ways the psychology community is more cacophony than chorus. It comprises several subgroups who view the world in very different and highly incompatible ways. Psychology's problems with replication don't require data. To produce conflicting conclusions, all you need is two psychologists. By adhering to different theoretical paradigms, psychologists can diametrically disagree on even the basic details of how human beings behave. Given this state of affairs, it is not difficult to see how diverging research findings might then also arise. In essence, psychology's replication crisis is bolstered by a fundamental paradigmatic crisis. It is to this problem of fragmentation that we now turn.

Index