BIAS ■ HIDDEN FLEXIBILITY ■ UNRELIABILITY ■ DATA HOARDING ■ CORRUPTIBILITY ■ INTERNMENT

THE 7 DEADLY SINS OF PSYCHOLOGY

A Manifesto *for* Reforming *the* Culture *of* Scientific Practice

CHRIS CHAMBERS

BIAS ■ BEAN COUNTING ■ INTERNMENT ■ CORRUPTIBILITY ■ DATA HOARDING ■ UNRELIABILITY

THE SEVEN DEADLY SINS OF PSYCHOLOGY

THE SEVEN DEADLY SINS OF PSYCHOLOGY

A Manifesto for Reforming the Culture of Scientific Practice

CHRIS CHAMBERS

Princeton University Press
Princeton & Oxford

CONTENTS

PREFACE

This book is borne out of what I can only describe as a deep personal frustration with the working culture of psychological science. I have always thought of our professional culture as a castle—a sanctuary of endeavor built long ago by our forebears. Like any home it needs constant care and attention, but instead of repairing it as we go we have allowed it to fall into a state of disrepair. The windows are dirty and opaque. The roof is leaking and won't keep out the rain for much longer. Monsters live in the dungeon.

Despite its many flaws, the castle has served me well. It sheltered me during my formative years as a junior researcher and advanced me to a position where I can now talk openly about the need for renovation. And I stress *renovation* because I am not suggesting we demolish our stronghold and start over. The foundations of psychology are solid, and the field has a proud legacy of discovery. Our founders—Helmholtz, Wundt, James—built it to last.

After spending fifteen years in psychology and its cousin, cognitive neuroscience, I have nevertheless reached an unsettling conclusion. If we continue as we are then psychology will diminish as a reputable science and could very well disappear. If we ignore the warning signs now, then in a hundred years or less, psychology may be regarded as one in a long line of quaint scholarly indulgences, much as we now regard alchemy or phrenology. Our descendants will smile tolerantly at this pocket of academic antiquity, nod sagely to one another about the protoscience that was psychology, and conclude that we were subject to the "limitations of the time." Of course, few sciences are likely to withstand the judgment of history, but it is by our research practices rather than our discoveries that psychology will be judged most harshly. And that judgment will be this: like so many other "soft" sciences, we found ourselves trapped within a culture where the *appearance* of science was seen as an appropriate replacement for the *practice* of science.

In this book I'm going to show how this distortion penetrates many aspects of our professional lives as scientists. The journey will be grim in

places. Using the seven deadly sins as a metaphor, I will explain how unchecked bias fools us into seeing what we want to see; how we have turned our backs on fundamental principles of the scientific method; how we treat the data we acquire as personal property rather than a public resource; how we permit academic fraud to cause untold damage to the most vulnerable members of our community; how we waste public resources on outdated forms of publishing; and how, in assessing the value of science and scientists, we have surrendered expert judgment to superficial bean counting. I will hope to convince you that in the quest for genuine understanding, we must be unflinching in recognizing these failings and relentless in fixing them.

Within each chapter, and in a separate final chapter, I will recommend various reforms that highlight two core aspects of science: transparency and reproducibility. To survive in the twenty-first century and beyond we must transform our secretive and fragile culture into a truly open and rigorous science—one that celebrates openness as much as it appreciates innovation, that prizes robustness as much as novelty. We must recognize that the old way of doing things is no longer fit for purpose and find a new path.

At its broadest level this book is intended for anyone who is interested in the practice and culture of science. Even those with no specific interest in psychology have reasons to care about the problems we face. Malpractice in any field wastes precious public funding by pursuing lines of enquiry that may turn out to be misleading or bogus. For example, by suppressing certain types of results from the published record, we risk introducing ineffective clinical treatments for mental health conditions such as depression and schizophrenia. In the UK, where the socioeconomic impact of research is measured as part of a regular national exercise called the Research Excellence Framework (REF), psychology has also been shown to influence a wide range of real-world applications. The 2014 REF reported over 450 "impact case studies" where psychological research has shaped public policy or practice, including (to name just a few) the design and uptake of electric cars, strategies for minimizing exam anxiety, the development of improved police interviewing techniques that account for the limits of human memory, setting of urban speed limits based on discoveries in vision science, human factors that are important for effective space exploration, government strategies for dealing with climate change that take into account public perception of risk, and plain packaging of tobacco products.[1] From its

most basic roots to its most applied branches, psychology is a rich part of public life and a key to understanding many global problems; therefore the deadly sins discussed here are a problem for society as a whole.

Some of the content, particularly sections on statistical methods, will be most relevant to the recently embarked researcher—the undergraduate student, PhD student, or early-career scientist—but there are also important messages throughout the book for more senior academics who manage their own laboratories or institutions, and many issues are also relevant to journalists and science writers. To aid the accessibility of source material for different audiences I have referred as much as possible to open access literature. For articles that are not open access, a Google Scholar search of the article title will often reveal a freely available electronic copy. I have also drawn on more contemporary forms of communication, including freely available blog entries and social media.

I owe a great debt to many friends, academic colleagues, journal editors, science writers, journalists, press officers, and policy experts, for years of inspiration, critical discussions, arguments, and in some cases interviews that fed into this work, including: Rachel Adams, Chris Allen, Micah Allen, Adam Aron, Vaughan Bell, Sven Bestmann, Ananyo Bhattacharya, Dorothy Bishop, Fred Boy, Todd Braver, Björn Brembs, Jon Brock, Jon Butterworth, Kate Button, Iain Chalmers, David Colquhoun, Molly Crockett, Stephen Curry, Helen Czerski, Zoltan Dienes, the late Jon Driver, Malte Elson, Alex Etz, John Evans, Eva Feredoes, Matt Field, Agneta Fischer, Birte Forstmann, Fiona Fox, Andrew Gelman, Tom Hardwicke, Chris Hartgerink, Tom Hartley, Mark Haselgrove, Steven Hill, Alex Holcombe, Aidan Horner, Macartan Humphreys, Hans Ijzerman, Helen Jamieson, Alok Jha, Gabi Jiga-Boy, Ben Johnson, Rogier Kievit, James Kilner, Daniël Lakens, Natalia Lawrence, Keith Laws, Katie Mack, Leah Maizey, Jason Mattingley, Rob McIntosh, Susan Michie, Candice Morey, Richard Morey, Simon Moss, Ross Mounce, Nils Mulhert, Kevin Murphy, Suresh Muthukumaraswamy, Bas Neggers, Neuroskeptic, Kia Nobre, Dave Nussbaum, Hans Op de Beeck, Ivan Oransky, Damian Pattinson, Andrew Przybylski, James Randerson, Geraint Rees, Ged Ridgway, Robert Rosenthal, Pia Rotshtein, Jeff Rouder, Elena Rusconi, Adam Rutherford, Chris Said, Ayse Saygin, Anne Scheel, Sam Schwarzkopf, Sophie Scott, Dan Simons, Jon Simons, Uri Simonsohn, Sanjay Srivastava, Mark Stokes, Petroc Sumner, Mike Taylor, Jon Tennant, Eric Turner, Carien van Reekum, Simine Vazire, Essi Viding, Solveiga Vivian-Griffiths, Matt

Wall, Tony Weidberg, Robert West, Jelte Wicherts, Ed Wilding, Andrew Wilson, Tal Yarkoni, Ed Yong, and Rolf Zwaan. Sincere thanks go to Sergio Della Sala and Toby Charkin for their collaboration and fortitude in championing Registered Reports at *Cortex*, Brian Nosek, David Mellor, and Sara Bowman for providing Registered Reports with such a welcoming home at the Center for Open Science, and to the Royal Society, particularly publisher Phil Hurst and publishing director Stuart Taylor, for embracing Registered Reports long before any other multidisciplinary journal. I also owe a debt of gratitude to Marcus Munafò for joining me in promoting Registered Reports at every turn, and to the 83 scientists who signed our *Guardian* open letter calling for installation of the format within all life science journals. Finally, I extend a special thanks to Dorothy Bishop, the late (and much missed) Alex Danchev, Dee Danchev, Zoltan Dienes, Pete Etchells, Hal Pashler, Frederick Verbruggen, and E. J. Wagenmakers for extensive draft reading and discussion, to Anastasiya Tarasenko for creating the chapter illustrations, and to my editors Sarah Caro and Eric Schwartz for their patience and sage advice throughout this journey.

THE SEVEN DEADLY SINS OF PSYCHOLOGY

CHAPTER 1

The Sin of Bias

The human understanding when it has once adopted
an opinion . . . draws all things else to support
and agree with it.
—Francis Bacon, 1620

History may look back on 2011 as the year that changed psychology forever. It all began when the *Journal of Personality and Social Psychology* published an article called "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect."[1] The paper, written by Daryl Bem of Cornell University, reported a series of experiments on *psi* or "precognition," a supernatural phenomenon that supposedly enables people to see events in the future. Bem, himself a reputable psychologist, took an innovative approach to studying *psi*. Instead of using discredited parapsychological methods such as card tasks or dice tests, he selected a series of gold-standard psychological techniques and modified them in clever ways.

One such method was a reversed priming task. In a typical priming task, people decide whether a picture shown on a computer screen is linked to a positive or negative emotion. So, for example, the participant might decide whether a picture of kittens is pleasant or unpleasant. If a word that "primes" the same emotion is presented immediately before the picture (such as the word "joy" followed by the picture of kittens), then people find it easier to judge the emotion of the picture, and they respond faster. But if the prime and target trigger opposite emotions then the task becomes more difficult because the emotions conflict (e.g., the word "murder" followed by kittens). To test for the existence of precognition, Bem reversed the order of this experiment and found that primes delivered *after* people had responded seemed to influence their reaction times. He also reported similar "retroactive" effects on memory. In one of his experiments, people were overall better at recalling specific words from a list that were also included in a practice task, with the catch that the so-called practice was undertaken *after* the recall task rather than before. On this basis, Bem argued that the participants were able to benefit in the past from practice they had completed in the future.

As you might expect, Bem's results generated a flood of confusion and controversy. How could an event in the future possibly influence someone's reaction time or memory in the past? If precognition truly did exist, in even a tiny minority of the population, how is it that casinos or stock markets turn profits? And how could such a bizarre conclusion find a home in a reputable scientific journal?

Scrutiny at first turned to Bem's experimental procedures. Perhaps there was some flaw in the methods that could explain his results, such as failing to randomize the order of events, or some other subtle experimental error. But these aspects of the experiment seemed to pass muster, leaving the research community facing a dilemma. If true, precognition would be the most sensational discovery in modern science. We would have to accept the existence of time travel and reshape our entire understanding of cause and effect. But if false, Bem's results would instead point to deep flaws in standard research practices—after all, if accepted practices could generate such nonsensical findings, how can any published findings in psychology be trusted? And so psychologists faced an unenviable choice between, on the one hand, accepting an impossible scientific conclusion and, on the other hand, swallowing an unpalatable professional reality.

The scientific community was instinctively skeptical of Bem's conclusions. Responding to a preprint of the article that appeared in late 2010, the psychologist Joachim Krueger said: "My personal view is that this is ridiculous and can't be true."[2] After all, extraordinary claims require extraordinary evidence, and despite being published in a prestigious journal, the statistical strength of Bem's evidence was considered far from extraordinary.

Bem himself realized that his results defied explanation and stressed the need for independent researchers to replicate his findings. Yet doing so proved more challenging than you might imagine. One replication attempt by Chris French and Stuart Ritchie showed no evidence whatsoever of precognition but was rejected by the same journal that published Bem's paper. In this case the journal didn't even bother to peer review French and Ritchie's paper before rejecting it, explaining that it "does not publish replication studies, whether successful or unsuccessful."[3] This decision may sound bizarre, but, as we will see, contempt for replication is common in psychology compared with more established sciences. The most prominent psychology journals selectively publish findings that they consider to be original, novel, neat, and above all positive. This *publication bias*, also known as the "file-drawer effect," means that studies that fail to show statistically significant effects, or that reproduce the work of others, have such low priority that they are effectively censored from the scientific record. They either end up in the file drawer or are never conducted in the first place.

Publication bias is one form of what is arguably the most powerful fallacy in human reasoning: *confirmation bias*. When we fall prey to confirmation bias, we seek out and favor evidence that agrees with our existing beliefs, while at the same time ignoring or devaluing evidence that doesn't. Confirmation bias corrupts psychological science in several ways. In its simplest form, it favors the publication of positive results—that is, hypothesis tests that reveal statistically significant differences or associations between conditions (e.g., A is greater than B; A is related to B, vs. A is the same as B; A is unrelated to B). More insidiously, it contrives a measure of scientific reproducibility in which it is possible to replicate but never falsify previous findings, and it encourages altering the hypotheses of experiments after the fact to "predict" unexpected outcomes. One of the most troubling aspects of psychology is that the academic community has refused to unanimously condemn such behavior. On the contrary, many psychologists acquiesce to these practices and even embrace them as survival skills in a culture where researchers must *publish or perish*.

Within months of appearing in a top academic journal, Bem's claims about precognition were having a powerful, albeit unintended, effect on the psychological community. Established methods and accepted publishing practices fell under renewed scrutiny for producing results that appear convincing but are almost certainly false. As psychologist Eric-Jan Wagenmakers and colleagues noted in a statistical demolition of Bem's paper: "Our assessment suggests that something is deeply wrong with the way experimental psychologists design their studies and report their statistical results."[4] With these words, the storm had broken.

A Brief History of the "Yes Man"

To understand the different ways that bias influences psychological science, we need to take a step back and consider the historical origins and basic research on confirmation bias. Philosophers and scholars have long recognized the "yes man" of human reasoning. As early as the fifth century BC, the historian Thucydides noted words to the effect that "[w]hen a man finds a conclusion agreeable, he accepts it without argument, but when he finds it disagreeable, he will bring against it all the forces of logic and reason." Similar sentiments were echoed by Dante, Bacon, and Tolstoy. By the mid-

twentieth century, the question had evolved from one of philosophy to one of science, as psychologists devised ways to measure confirmation bias in controlled laboratory experiments.

Since the mid-1950s, a convergence of studies has suggested that when people are faced with a set of observations (data) and a possible explanation (hypothesis), they favor tests of the hypothesis that seek to confirm it rather than falsify it. Formally, what this means is that people are biased toward estimating the probability of data if a particular hypothesis is *true*, *p*(data|hypothesis) rather than the opposite probability of it being false, *p*(data|~hypothesis). In other words, people prefer to ask questions to which the answer is "yes," ignoring the maxim of philosopher Georg Henrik von Wright that "no confirming instance of a law is a verifying instance, but . . . any disconfirming instance is a falsifying instance."[5]

Psychologist Peter Wason was one of the first researchers to provide laboratory evidence of confirmation bias. In one of several innovative experiments conducted in the 1960s and 1970s, he gave participants a sequence of numbers, such as 2-4-6, and asked them to figure out the rule that produced it (in this case: *three numbers in increasing order of magnitude*).[6] Having formed a hypothesis, participants were then allowed to write down their own sequence, after which they were told whether their sequence was consistent or inconsistent with the actual rule. Wason found that participants showed a strong bias to test various hypotheses by *confirming* them, even when the outcome of doing so failed to eliminate plausible alternatives (such as *three even numbers*). Wason's participants used this strategy despite being told in advance that "your aim is not simply to find numbers which conform to the rule, but to discover the rule itself."

Since then, many studies have explored the basis of confirmation bias in a range of laboratory-controlled situations. Perhaps the most famous of these is the ingenious Selection Task, which was also developed by Wason in 1968.[7] The Selection Task works like this. Suppose I were to show you four cards on a table, labeled D, B, 3, and 7 (see figure 1.1). I tell you that if the card shows a letter on one side then it will have a number on the other side, and I provide you with a more specific rule (hypothesis) that may be true or false: "*If there is a D on one side of any card, then there is a 3 on its other side*." Finally, I ask you to tell me which cards you would need to turn over in order to determine whether this rule is true or false. Leaving an informative card unturned or turning over an uninformative card (i.e., one that

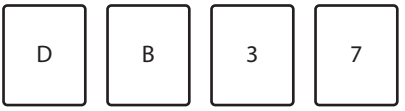


FIGURE 1.1. Peter Wason's Selection Task for measuring confirmation bias. Four cards are placed face down on a table. You're told that if there is letter on one side then there will always be a number on the other side. Then you are given a specific hypothesis: *If there is a D on one side then there is a 3 on its other side*. Which cards would you turn over to test whether this hypothesis is true or false?

doesn't test the rule) would be considered an incorrect response. Before reading further, take a moment and ask yourself, which cards would you choose and which would you avoid?

If you chose D and avoided B then you're in good company. Both responses are correct and are made by the majority of participants. Selecting D seeks to test the rule by confirming it, whereas avoiding B is correct because the flip side would be uninformative regardless of the outcome.

Did you choose 3? Wason found that most participants did, even though 3 should be avoided. This is because if the flip side *isn't* a D, we learn nothing—the rule states that cards with D on one side are paired a 3 on the other, not that D is the *only* letter to be paired with a 3 (drawing such a conclusion would be a logical fallacy known as "affirming the consequent"). And even if the flip side *is* a D then the outcome would be consistent with the rule but wouldn't confirm it, for exactly the same reason.

Finally, did you choose 7 or avoid it? Interestingly, Wason found that few participants selected 7, even though doing so is correct—in fact, it is just as correct as selecting D. If the flip side to 7 were discovered to be a D then the rule would be categorically disproven—a logical test of what's known as the "contrapositive." And herein lies the key result: the fact that most participants correctly select D but fail to select 7 provides evidence that people seek to test rules or hypotheses by confirming them rather than by falsifying them.

Wason's findings provided the first laboratory-controlled evidence of confirmation bias, but centuries of informal observations already pointed strongly to its existence. In a landmark review, psychologist Raymond Nickerson noted how confirmation bias dominated in the witchcraft trials of the middle ages.[8] Many of these proceedings were a foregone conclusion, seeking only to obtain evidence that confirmed the guilt of the accused. For

instance, to test whether a person was a witch, the suspect would often be plunged into water with stones tied to her feet. If she rose then she would be proven a witch and burned at the stake. If she drowned then she was usually considered innocent or a witch of lesser power. Either way, being suspected of witchcraft was tantamount to a death sentence within a legal framework that sought only to confirm accusations. Similar biases are apparent in many aspects of modern life. Popular TV programs such as *CSI* fuel the impression that forensic science is bias-free and infallible, but in reality the field is plagued by confirmation bias.[9] Even at the most highly regarded agencies in the world, forensic examiners can be biased toward interpreting evidence that confirms existing suspicions. Doing so can lead to wrongful convictions, even when evidence is based on harder data such as fingerprints and DNA tests.

Confirmation bias also crops up in the world of science communication. For many years it was assumed that the key to more effective public communication of science was to fill the public's lack of knowledge with facts—the so-called deficit model.[10] More recently, however, this idea has been discredited because it fails to take into account the prior beliefs of the audience. The extent to which we assimilate new information about popular issues such as climate change, vaccines, or genetically modified foods is susceptible to a confirmation bias in which evidence that is consistent with our preconceptions is favored, while evidence that flies in the face of them is ignored or attacked. Because of this bias, simply handing people more facts doesn't lead to more rational beliefs. The same problem is reflected in politics. In his landmark 2012 book, the *Geek Manifesto*, Mark Henderson laments the cherry-picking of evidence by politicians in order to reinforce a predetermined agenda. The resulting "policy-based evidence" is a perfect example of confirmation bias in practice and represents the antithesis of how science should be used in the formulation of evidence-based policy.

If confirmation bias is so irrational and counterproductive, then why does it exist? Many different explanations have been suggested based on cognitive or motivational factors. Some researchers have argued that it reflects a fundamental limit of human cognition. According to this view, the fact that we have incomplete information about the world forces us to rely on the memories that are most easily retrieved (the so-called availability heuristic), and this reliance could fuel a bias toward what we think we already know. On the other hand, others have argued that confirmation bias

is the consequence of an innate "positive-test strategy"—a term coined in 1987 by psychologists Joshua Klayman and Young-Won Ha.[11] We already know that people find it easier to judge whether a positive statement is true or false (e.g., "there are apples in the basket") compared to a negative one ("there are no apples in the basket"). Because judgments of presence are easier than judgments of absence, it could be that we prefer positive tests of reality over negative ones. By taking the easy road, this bias toward positive thoughts could lead us to wrongly accept evidence that agrees positively with our prior beliefs.

Against this backdrop of explanations for why an irrational bias is so pervasive, psychologists Hugo Mercier and Dan Sperber have suggested that confirmation bias is in fact perfectly rational in a society where winning arguments is more important than establishing truths.[12] Throughout our upbringing, we are taught to defend and justify the beliefs we hold, and less so to challenge them. By interpreting new information according to our existing preconceptions we boost our self-confidence and can argue more convincingly, which in turn increases our chances of being regarded as powerful and socially persuasive. This observation leads us to an obvious proposition: If human society is constructed so as to reward the act of winning rather than being correct, who would be surprised to find such incentives mirrored in scientific practices?

Neophilia: When the Positive and New Trumps the Negative but True

The core of any research psychologist's career—and indeed many scientists in general—is the rate at which they publish empirical articles in high-quality peer-reviewed journals. Since the peer-review process is competitive (and sometimes extremely so), publishing in the most prominent journals equates to a form of "winning" in the academic game of life.

Journal editors and reviewers assess submitted manuscripts on many grounds. They look for flaws in the experimental logic, the research methodology, and the analyses. They study the introduction to determine whether the hypotheses are appropriately grounded in previous research. They scrutinize the discussion to decide whether the paper's conclusions are justified by the evidence. But reviewers do more than merely critique

the rationale, methodology, and interpretation of a paper. They also study the results themselves. How important are they? How exciting? How much have we learned from this study? Is it a breakthrough? One of the central (and as we will see, lamentable) truths in psychology is that exciting positive results are a key factor in publishing—and often a requirement. The message to researchers is simple: if you want to win in academia, publish as many papers as possible in which you provide positive, novel results.

What does it mean to find "positive" results? Positivity in this context doesn't mean that the results are uplifting or good news—it refers to whether the researchers found a reliable difference in measurements, or a reliable relationship, between two or more study variables. For example, suppose you wanted to test the effect of a cognitive training intervention on the success of dieting in people trying to lose weight. First you conduct a literature review, and, based on previous studies, you decide that boosting people's self-control might help. Armed with a good understanding of existing work, you design a study that includes two groups. The experimental group perform a computer task in which they are trained to respond to images of foods, but crucially, to refrain from responding to images of particular junk foods. They perform this task every day for six weeks, and you measure how much weight they lose by the end of the experiment. The control group does a similar task with the same images but responds to all of them—and you measure weight loss in that group as well.

The null hypothesis (called "H0") in this case is that there should be no difference in weight loss—your training intervention has no effect on whether people gain or lose weight. The alternative hypothesis (called "H1") is that the training intervention should boost people's ability to refrain from eating junk foods, and so the amount of weight loss should be greater in the treatment group compared with the control group. A positive result would be finding a statistically significant difference in weight loss between the groups (or in technical terms, "rejecting H0"), and a negative result would be failing to show any significant difference (or in other words, "failing to reject H0"). Note how I use the term "failing." This language is key because, in our current academic culture, journals indeed regard such outcomes as scientific failures. Regardless of the fact that the rationale and methods are identical in each outcome, psychologists find negative results much harder to publish than positive results. This is because positive results are regarded by journals as reflecting a greater degree of scientific advance and interest

to readers. As one journal editor said to me, "Some results are just more interesting and important than others. If I do a randomized trial on a novel intervention based on a long-shot and find no effect that is not a great leap forward. However, if the same study shows a huge benefit that is a more important finding."

This publication bias toward positive results also arises because of the nature of conventional statistical analyses in psychology. Using standard methods developed by Neyman and Pearson, positive results reject H0 in favor of the alternative hypothesis (H1). This statistical approach—called null hypothesis significance testing—estimates the probability (p) of an effect of the same or greater size being obtained if the null hypothesis were true. Crucially, it doesn't estimate the probability of the null hypothesis *itself* being true: p values estimate the probability of a given effect or more extreme arising given the hypothesis, rather than the probability of a particular hypothesis given the effect. This means that while a statistically significant result (by convention, $p<.05$) allows the researcher to *reject* H0, a statistically nonsignificant result ($p>.05$) doesn't allow the researcher to *accept* H0. All the researcher can conclude from a statistically nonsignificant outcome is that H0 might be true, or that the data might be insensitive. The interpretation of statistically nonsignificant effects is therefore inherently inconclusive.

Consider the thought process this creates in the minds of researchers. If we can't test directly whether there is *no* difference between experimental conditions, then it makes little sense to design an experiment in which the null hypothesis would ever be the focus of interest. Instead, psychologists are trained to design experiments in which findings of interest would always be *positive*. This bias in experimental design, in turn, means that students in psychology enter their research careers reciting the mantra "Never predict the null hypothesis." If researchers can never predict the null hypothesis, and if positive results are considered more interesting to journals than negative results, then the inevitable outcome is a bias in which the peer-reviewed literature is dominated by positive findings that reject H0 in favor of H1, and in which most of the negative or nonsignificant results remain unpublished. To ensure that they keep winning in the academic game, researchers are thus pushed into finding positive results that agree with their expectations—a mechanism that incentivizes and rewards confirmation bias.

All this might sound possible in theory, but is it true? Psychologists have known since the 1950s that journals are predisposed toward publishing positive results, but, historically, it has been difficult to quantify how much publication bias there really is in psychology.[13] One of the most compelling analyses was reported in 2010 by psychologist Daniele Fanelli from the University of Edinburgh.[14] Fanelli reasoned, as above, that any domain of the scientific literature that suffers from publication bias should be dominated by positive results that support the stated hypothesis (H1). To test this idea, he collected a random sample of more than 2,000 published journal articles from across the full spectrum of science, ranging from the space sciences to physics and chemistry, through to biology, psychology, and psychiatry. The results were striking. Across all sciences, positive outcomes were more common than negative ones. Even for space science, which published the highest percentage of negative findings, 70 percent of the sampled articles supported the stated hypothesis. Crucially, this bias was highest in psychology, topping out at 91 percent. It is ironic that psychology—the discipline that produced the first empirical evidence of confirmation bias—is at the same time one of the most vulnerable *to* confirmation bias.

The drive to publish positive results is a key cause of publication bias, but it still explains only half the problem. The other half is the quest for novelty. To compete for publication at many journals, articles must either adopt a novel methodology or produce a novel finding—and preferably both. Most journals that publish psychological research judge the merit of manuscripts, in part, according to novelty. Some even refer explicitly to novelty as a policy for publication. The journal *Nature* states that to be considered for peer review, results must be "novel" and "arresting,"[15] while the journal *Cortex* notes that empirical Research Reports must "report important and novel material."[16] The journal *Brain* warns authors that "some [manuscripts] are rejected without peer review owing to lack of novelty,"[17] and *Cerebral Cortex* goes one step further, noting that even after peer review, "final acceptance of papers depends not just on technical merit, but also on subjective ratings of novelty."[18] Within psychology proper, *Psychological Science*, a journal that claims to be the highest-ranked in psychology, prioritizes papers that produce "breathtaking" findings.[19]

At this point, you might well ask: what's wrong with novelty? After all, in order for something to be marked as discovered, surely it can't have been observed already (so it must be a novel result), and isn't it also reasonable

to assume that researchers seeking to produce novel results might need to adopt new methods? In other words, by valuing novelty aren't journals simply valuing discovery? The problem with this argument is the underlying assumption that every observation in psychological research can be called a discovery—that every paper reports a clear and definitive fact. As with all scientific disciplines, this is far from the truth. Most research findings in psychology are probabilistic rather than deterministic: conventional statistical tests talk to us in terms of probabilities rather than proofs. This in turn means that no single study and no one paper can lay claim to a discovery. Discovery depends wholly and without exception on the extent to which the original results can be repeated or *replicated* by other scientists, and not just once but over and over again. For example, it would not be enough to report only once that a particular cognitive therapy was effective at reducing depression; the result would need to be repeated many times in different groups of patients, and by different groups of researchers, for it be widely adopted as a public health intervention. Once a result has been replicated a satisfactory number of times using the same experimental method, it can then be considered *replicable* and, in combination with other replicable evidence, can contribute meaningfully to the theoretical or applied framework in which it resides. Over time, this mass accumulation of replicable evidence within different fields can allow theories to become accepted through consensus and in some cases can even become laws.

In science, prioritizing novelty hinders rather than helps discovery because it dismisses the value of direct (or close) replication. As we have seen, journals are the gatekeepers to an academic career, so if they value findings that are positive and novel, why would scientists ever attempt to replicate each other? Under a neophilic incentive structure, direct replication is discarded as boring, uncreative, and lacking in intellectual prowess.

Yet even in a research system dominated by positive bias and neophilia, psychologists have retained some realization that reproducibility matters. So, in place of unattractive direct replication, the community has reached for an alternative form of validation in which one experiment can be said to replicate the key concept or theme of another by following a different (novel) experimental method—a process known as *conceptual* replication. On its face, this redefinition of replication appears to satisfy the need to validate previous findings while also preserving novelty. Unfortunately, all it really does is introduce an entirely new and pernicious form of confirmation bias.

Replicating Concepts Instead of Experiments

In early 2012, a professor of psychology at Yale University named John Bargh launched a stinging public attack on a group of researchers who failed to replicate one of his previous findings.[20] The study in question, published by Bargh and colleagues in 1996, reported that priming participants unconsciously to think about concepts related to elderly people (e.g., words such as "retired," "wrinkle," and "old") caused them to walk more slowly when leaving the lab at the end of the experiment.[21] Based on these findings, Bargh claimed that people are remarkably susceptible to automatic effects of being primed by social constructs.

Bargh's paper was an instant hit and to date has been cited more than 3,800 times. Within social psychology it spawned a whole generation of research on social priming, which has since been applied in a variety of different contexts. Because of the impact the paper achieved, it would be reasonable to expect that the central finding must have been replicated many times and confirmed as being sound. Appearances, however, can be deceiving.

Several researchers had reported failures to replicate Bargh's original study, but few of these nonreplications have been published, owing to the fact that journals (and reviewers) disapprove of negative findings and often refuse to publish direct replications. One such attempted replication in 2008 by Hal Pashler and colleagues from the University of California San Diego was never published in an academic journal and instead resides at an online repository called PsychFileDrawer.[22] Despite more than doubling the sample size reported in the original study, Pashler and his team found no evidence of such priming effects—if anything they found the opposite result.

Does this mean Bargh was wrong? Not necessarily. As psychologist Dan Simons from the University of Illinois has noted, failing to replicate an effect does not necessarily mean the original finding was in error.[23] Nonreplications can emerge by chance, can be due to subtle changes in experimental methods between studies, or can be caused by the poor methodology of the researchers attempting the replication. Thus, nonreplications are themselves subject to the same tests of replicability as the studies they seek to replicate.

Nevertheless, the failed replication by Pashler and colleagues—themselves an experienced research team—raised a question mark over the status of Bargh's original study and hinted at the existence of an invisible file drawer

of unpublished failed replications. In 2012, another of these attempted replications came to light when Stéphane Doyen and colleagues from the University of Cambridge and Université Libre de Bruxelles also failed to replicate the elderly priming effect.[24] Their article appeared prominently in the peer-reviewed journal *PLOS ONE*, one of the few outlets worldwide that explicitly renounces neophilia and publication bias. The ethos of *PLOS ONE* is to publish any methodologically sound scientific research, regardless of subjective judgments as to its perceived importance or originality. In their study, Doyen and colleagues not only failed to replicate Bargh's original finding but also provided an alternative explanation for the original effect—rather than being due to a priming manipulation, it was the experimenters themselves who unwittingly induced the participants to walk more slowly by behaving differently or even revealing the hypothesis.

The response from Bargh was swift and contemptuous. In a highly publicized blogpost at psychologytoday.com entitled "Nothing in Their Heads,"[25] he attacked not only Doyen and colleagues as "incompetent or ill-informed," but also science writer Ed Yong (who covered the story)[26] for engaging in "superficial online science journalism," and *PLOS ONE* as a journal that "quite obviously does not receive the usual high scientific journal standards of peer-review scrutiny." Amid a widespread backlash against Bargh, his blogpost was swiftly (and silently) deleted but not before igniting a fierce debate about the reliability of social priming research and the status of replication in psychology more generally.

Doyen's article, and the response it generated, didn't just question the authenticity of the elderly priming effect; it also exposed a crucial disagreement about the definition of replication. Some psychologists, including Bargh himself, claimed that the original 1996 study had been replicated at length, while others claimed that it had *never* been replicated. How is this possible?

The answer, it turned out, was that different researchers were defining replication differently. Those who argued that the elderly priming effect had never been replicated were referring to *direct* replications: studies that repeat the method of a previous experiment as exactly as possible in order to reproduce the finding. At the time of writing, Bargh's central finding has been directly replicated just twice, and in each case with only partial success. In the first attempt, published six years after the original study,[27] the researchers showed the same effect but only in a subgroup of participants who scored

high on self-consciousness. In the second attempt, published another four years later, a different group of authors showed that priming elderly concepts slowed walking only in participants who held positive attitudes about elderly people; those who harbored negative attitudes showed the opposite effect.[28] Whether these partial replications are themselves replicable is unknown, but as we will see in chapter 2, hidden flexibility in the choices researchers make when analyzing their data (particularly concerning subgroup analyses) can produce spurious differences where none truly exist.

In contrast, those who argued that the elderly priming effect had been replicated many times were referring to the notion of "conceptual replication": the idea that the *principle* of unconscious social priming demonstrated in Bargh's 1996 study has been extended and applied in many different contexts. In a later blog post at psychologytoday.com called "Priming Effects Replicate Just Fine, Thanks," Bargh referred to some of these conceptual replications in variety of social behaviors, including attitudes and stereotypes unrelated to the elderly.[29]

The logic of "conceptual replication" is that if an experiment shows evidence for a particular phenomenon, you can replicate it by using a different method that the experimenter believes measures the same class of phenomenon. Psychologist Rolf Zwaan argues that conceptual replication has a legitimate role in psychology (and indeed all sciences) to test the extent to which particular phenomena depend on specific laboratory conditions, and to determine whether they can be generalized to new contexts.[30] The current academic culture, however, has gone further than merely valuing conceptual replication—it has allowed it to usurp direct replication. As much as we all agree about the importance of converging evidence, should we be seeking it out at the expense of knowing whether the phenomenon being generalized exists in the first place?

A reliance on conceptual replication is dangerous for three reasons.[31] The first is the problem of subjectivity. A conceptual replication can hold only if the different methods used in two different studies are measuring the same phenomenon. For this to be the case, some evidence must exist that they are. Even if we meet this standard, this raises the question of how similar the methods must be for a study to qualify as being conceptually replicated. Who decides and by what criteria?

The second problem is that a reliance on conceptual replications risks findings becoming *unreplicated* in the future. To illustrate how this could

happen, suppose we have three researchers, Smith, Jones, and Brown, who publish three scientific papers in sequence. Smith publishes the first paper, showing evidence for a particular phenomenon. Jones then uses a different method to show evidence for a phenomenon that appears similar to the one that Smith discovered. The psychological community decide that the similarity crosses some subjective threshold and so conclude that Jones "conceptually replicates" Smith. Now enter Brown. Brown isn't convinced that Smith and Jones are measuring the same phenomenon and suspects they are in fact describing *different* phenomena. Brown obtains evidence suggesting that this is indeed the case. In this way, Smith's finding that was previously considered replicated by Jones now assumes the bizarre status of becoming *un*replicated.

Finally, conceptual replication fuels an obvious confirmation bias. When two studies draw similar conclusions using different methods, the second study can be said to conceptually replicate the first. But what if the second study draws a very different conclusion—would it be claimed to conceptually *falsify* the first study? Of course not. Believers of the original finding would immediately (and correctly) point to the multitude of differences in methodology to explain the different results. Conceptual replications thus force science down a one-way street in which it is possible to confirm but never disconfirm previous findings. Through a reliance on conceptual replication, psychology has found yet another way to become enslaved to confirmation bias.

Reinventing History

So far we have seen how confirmation bias influences psychological science in two ways: through the pressure to publish results that are novel and positive, and by ousting direct replication in favor of bias-prone conceptual replication. A third, and especially insidious, manifestation of confirmation bias can be found in the phenomenon of hindsight bias. Hindsight bias is a form of creeping determinism in which we fool ourselves (and others) into believing that an observation was expected even though it actually came as a surprise.

It may seem extraordinary that any scientific discipline should be vulnerable to a fallacy that attempts to reinvent history. Indeed, under the classic

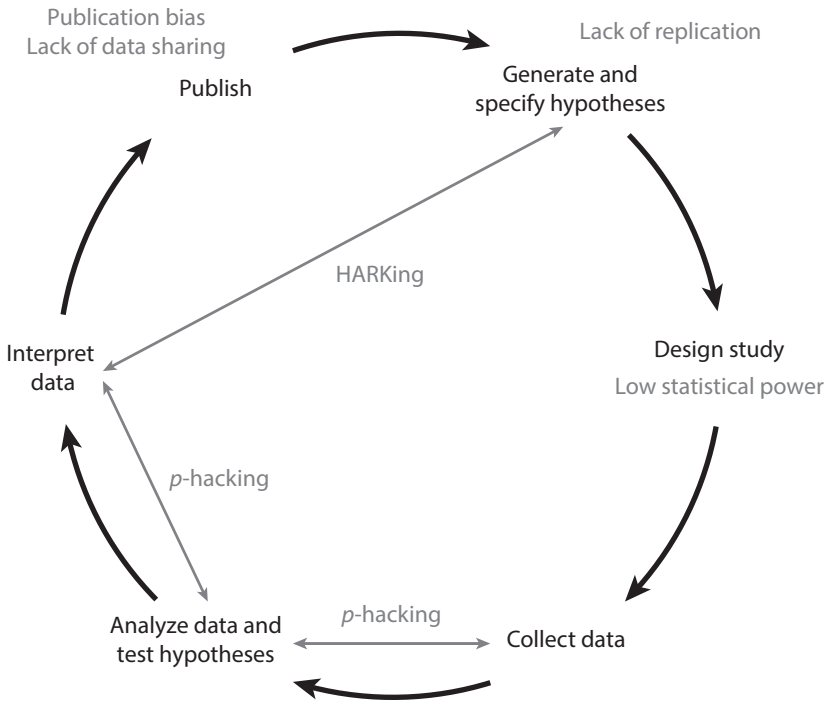


FIGURE 1.2. The hypothetico-deductive model of the scientific method is compromised by a range of questionable research practices. **Lack of replication** impedes the elimination of false discoveries and weakens the evidence base underpinning theory. **Low statistical power** (to be discussed in chapter 3) increases the chances of missing true discoveries and reduces the probability that obtained positive effects are real. Exploiting researcher degrees of freedom (***p*-hacking**—to be discussed in chapter 2) manifests in two general forms: collecting data until analyses return statistically significant effects, and selectively reporting analyses that reveal desirable outcomes. **HARKing**, or Hypothesizing After Results are Known, involves generating a hypothesis from the data and then presenting it as a priori. **Publication bias** occurs when journals reject manuscripts on the basis that they report negative or otherwise unattractive findings. Finally, **lack of data sharing** (to be discussed in chapter 4) prevents detailed meta-analysis and hinders the detection of data fabrication.

hypothetico-deductive (H-D) model of the scientific method, the research process is supposed to be protected against such bias (see figure 1.2). According to the H-D method, to which psychology at least nominally adheres, a scientist begins by formulating a hypothesis that addresses some aspect of a relevant theory. With the hypothesis decided, the scientist then conducts an experiment and allows the data to determine whether or not the

hypothesis was supported. This outcome then feeds into revision (and possible rejection) of the theory, stimulating an iterative cycle of hypothesis generation, hypothesis testing, and theoretical advance. A central feature of the H-D method is that the hypothesis is decided *before* the scientist collects and analyzes the data. By separating in time the prediction (hypothesis) from the estimate of reality (data), this method is designed to protect scientists from their own hindsight bias.

Unfortunately, much psychological research seems to pay little heed to this aspect of the scientific method. Since the hypothesis of an experiment is only rarely published in advance, researchers can covertly alter their predictions after the data have been analyzed in the interests of narrative flair. In psychology this practice is referred to as Hypothesizing After Results are Known (HARKing), a term coined in 1998 by psychologist Norbert Kerr.[32] HARKing is a form of academic deception in which the experimental hypothesis (H1) of a study is altered after analyzing the data in order to pretend that the authors predicted results that, in reality, were unexpected. By engaging in HARKing, authors are able to present results that seem neat and consistent with (at least some) existing research or their own previously published findings. This flexibility allows the research community to produce the kind of clean and confirmatory papers that psychology journals prefer while also maintaining the illusion that the research is hypothesis driven and thus consistent with the H-D method.

HARKing can take many forms, but one simple approach involves reversing the predictions after inspecting the data. Suppose that a researcher formulates the hypothesis that, based on the associations we form across our lifetime between the color red and various behavioral acts of stopping (e.g., traffic lights; stop signs; hazard signs), people should become more cautious in a gambling task when the stimuli used are red rather than white. After running the experiment, however, the researcher finds the opposite result: people gambled *more* when exposed to red stimuli. According to the H-D method, the correct approach here would be to report that the hypothesis was unsupported, admitting that additional experiments may be required to understand how this unexpected result arose and its theoretical implications. However, the researcher realizes that this conclusion may be difficult to publish without conducting those additional experiments, and he or she also knows that nobody reviewing the paper would be aware that

the original hypothesis was unsupported. So, to create a more compelling narrative, the researcher returns to the literature and searches for studies suggesting that being exposed to the color red can lead people to "see red," losing control and becoming *more* impulsive. Armed with a small number of cherry-picked findings, the researcher ignores the original (better grounded) rationale and rewrites the hypothesis to predict that people will actually gamble *more* when exposed to red stimuli. In the final published paper, the introduction section is written with this post hoc hypothesis presented as a priori.

Just how prevalent is this kind of HARKing? Norbert Kerr's survey of 156 psychologists in 1998 suggested that about 40 percent of respondents had observed HARKing by other researchers; strikingly, the surveyed psychologists also suspected that HARKing was about 20 percent more prevalent than the classic H-D method.[33] A more recent survey of 2,155 psychologists by Leslie John and colleagues estimated the true prevalence rate to be as high as 90 percent despite a self-admission rate of just 35 percent.[34]

Remarkably, not all psychologists agree that HARKing is a problem. Nearly 25 years before suggesting the existence of precognition, Daryl Bem claimed that if data are "strong enough" then researchers are justified in "subordinating or even ignoring [their] original hypotheses."[35] In other words, Bem argued that it is legitimate to subvert the H-D method, and to do so covertly, in order to preserve the narrative structure of a scientific paper.

Norbert Kerr and others have objected to this point of view, as well they might. First and foremost, because HARKing relies on deception, it violates the fundamental ethical principle that research should be reported honestly and completely. Deliberate HARKing may therefore lie on the same continuum of malpractice as research fraud. Secondly, the act of deception in HARKing leads the reader to believe that an obtained finding was more expected, and hence more reliable, than it truly is—this, in turn, risks distorting the scientific record to place undue certainty in particular findings and theories. Finally, in cases where a post hoc hypothesis is pitted against an alternative account that the author already knows was unsupported, HARKing creates the illusion of competitive hypothesis testing. Since a HARKed hypothesis can, by definition, never be disconfirmed, this contrived scenario further exacerbates confirmation bias.

The Battle against Bias

If confirmation bias is so much a part of human nature then what hope can we have of defeating it in science? In an academic culture that prizes novel results that confirm our expectations, is there any real chance of reform? We have known about the various manifestations of bias in psychology since the 1950s—and have done little to counteract them—so it is easy to see why many psychologists are cynical about the prospect of change. However, the tide is turning. Chapter 8 will address the set of changes we must make—and are already launching—to protect psychological science against bias and the other "deadly sins" that have become part of our academic landscape. Some of these reforms are already bearing fruit.

Our starting point for any program of reform must be the acceptance that we can never completely eliminate confirmation bias—in Nietzsche's words we are *human, all too human*. Decades of psychological research shows how bias is woven into the fabric of cognition and, in many situations, operates unconsciously. So, rather than waging a fruitless war on our own nature, we would do better to accept imperfection and implement measures that protect the outcome of science as much as possible from our inherent flaws as human practitioners.

One such protection against bias is study preregistration. We will return to the details of preregistration in chapter 8, but for now it is useful to consider how publicly registering our research intentions *before* we collect data can help neutralize bias. Consider the three main manifestations of confirmation bias in psychology: publication bias, conceptual replication, and HARKing. In each case, a strong motivation for engaging in these practices is not to generate high-quality, replicable science, but to produce results that are publishable and perceived to be of interest to other scientists. Journals enforce publication bias because they believe that novel, positive results are more likely to indicate discoveries that their readers will want to see; by comparison, replications and negative findings are considered boring and relatively lacking in intellectual merit. To fit with the demands of journals, psychologists have thus replaced direct replication with conceptual replication, maintaining the comfortable but futile delusion that our science values replication while still satisfying the demands of novelty and originality. Finally, as we have seen, many researchers engage in HARKing because

they realize that failing to confirm their own hypothesis is regarded as a form of intellectual failure.

Study preregistration helps overcome these problems by changing the incentive structure to value "good science" over and above "good results." The essence of preregistration is that the study rationale, hypotheses, experimental methods, and analysis plan are stated publicly in advance of collecting data. When this process is undertaken through a peer-reviewed journal, it forces journal editors to make publishing decisions before results exist. This, in turn, prevents publication bias by ensuring that whether results are positive or negative, novel or familiar, groundbreaking or incremental, is irrelevant to whether the science will be published. Similarly, since authors will have stated their hypotheses in advance, preregistration prevents HARKing and ensures adherence to the H-D model of the scientific method. As we will see in chapter 2, preregistration also prevents researchers from cherry-picking results that they believe generate a desirable narrative.

In addition to study preregistration, bias can be reduced by reforming statistical practice. As discussed earlier, one reason negative findings are regarded as less interesting is our cultural reliance on null hypothesis significance testing (NHST). NHST can only ever tell us whether the null hypothesis is rejected, and never whether it is supported. Our reliance on this one-sided statistical approach inherently places greater weight on positive findings. However, by shifting to alternative Bayesian statistical methods, we can test all potential hypotheses (H0, H1 . . . Hn) fairly as legitimate possible outcomes. We will explore this alternative method in more detail in chapter 3.

As we take this journey it is crucial that individual scientists from every level feel empowered to promote reform without damaging their careers. Confirmation bias is closely allied with "groupthink"—a pernicious social phenomenon in which a consensus of behavior is mistaken for a convergence of informed evidence. The herd doesn't always make the most rational or intelligent decisions, and groupthink can stifle innovation and critical reflection. To ensure the future of psychological science, it is incumbent on us as psychologists to recognize and challenge our own biases.

CHAPTER 2

The Sin of Hidden Flexibility

Torture numbers and they will confess to anything.
—Gregg Easterbrook, 1999

In 2008, British illusionist Derren Brown presented a TV program called *The System* in which he claimed he could predict, with certainty, which horse would win at the racetrack. The show follows Khadisha, a member of the public, as Brown provides her with tips on upcoming races. In each case the tips pay off, and after succeeding five times in a row Khadisha decides to bet as much money as possible on a sixth and final race. The twist in the program is that Brown has no system—Khadisha is benefiting from nothing more than chance. Unknown to her until after placing her final bet, Brown initially recruited 7,776 members of the public and provided each of them with a unique combination of potential winners. Participants with a losing horse were successively eliminated at each of six stages, eventually leaving just one participant who had won every time—and that person just happened to be Khadisha. By presenting the story from Khadisha's perspective, Brown created the illusion that her winning streak was too unlikely to be random—and so must be caused by The System—when in fact it was explained entirely by chance.

Unfortunately for science, the hidden flexibility that Brown used to generate false belief in The System is the same mechanism that psychologists exploit to produce results that are attractive and easy to publish. Faced with the career pressure to publish positive findings in the most prestigious and selective journals, it is now standard practice for researchers to analyze complex data in many different ways and report only the most interesting and statistically significant outcomes. Doing so deceives the audience into believing that such outcomes are credible, rather than existing within an ocean of unreported negative or inconclusive findings. Any conclusions drawn from such tests will, at best, overestimate of the size of any real effect. At worst they could be entirely false.

By torturing numbers until they produce publishable outcomes, psychology commits our second mortal sin: that of exploiting hidden analytic flexibility. Formally, hidden flexibility is one manifestation of the "fallacy of incomplete evidence," which arises when we frame an argument without taking into account the full set of information available. Although hidden flexibility is itself a form of research bias, its deep and ubiquitous nature in psychology earns it a dedicated place in our hall of shame.

p-Hacking

As we saw earlier, the dominant approach for statistical analysis in psychological science is a set of techniques called null hypothesis significance testing (NHST). NHST estimates the probability of an obtained positive effect, or one greater, being observed in a set of data if the null hypothesis (H0) is true and no effect truly exists. Importantly, the *p* value doesn't tell us the probability of H0 itself being true, and it doesn't indicate the size or reliability of the obtained effect—instead what it tells us is how surprised we should be to obtain the current effect, or one more extreme, if H0 *were* to be true.[1] The smaller the *p* value, the greater our surprise would be and the more confidently we can reject H0.

Since the 1920s, the convention in psychology has been to require a *p* value of less than .05 in order to categorically reject H0. This significance threshold is known as α—the probability of falsely declaring a positive effect when, in fact, there isn't one. Under NHST, a false positive or Type I error occurs when we incorrectly reject a true H0. The α threshold thus indicates the maximum allowable probability of a Type I error in order to reject H0 and conclude that a statistically significant effect is present.

Why is α set to .05, you might ask? The .05 convention is arbitrary, as noted by Ronald Fisher—one of the architects of NHST—nearly a century ago:

> If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level.[2]

Setting the α threshold to .05 theoretically allows up to 1 in 20 false rejections of H0 across a set of independent significance tests. Some have argued that this threshold is too liberal and leads to a scientific literature built on weak findings that are unlikely to replicate.[3] Furthermore, even if we believe that it is acceptable for 5 percent of statistically significant results to be false positives, the truth is that exploiting analytic flexibility increases α even more, increasing the *actual* rate of false positives.

This flexibility arises because researchers make analytic decisions after inspecting their data and are faced with many analysis options that can be

considered defensible yet produce slightly different *p* values. For instance, given a distribution of reaction time values, authors have the option of excluding statistical outliers (such as very slow responses) within each participant. They also have the option of excluding entire participants on the same basis. If they decide to adopt either or both of these approaches, there are then many available methods they could use, each of which could produce slightly different results. As well as being flexible, a key feature of such decisions is that they are hidden and never published. The rules of engagement do not require authors to specify which analytic decisions were a priori (confirmatory) and which were post hoc (exploratory)—in fact, such transparency is likely to penalize authors competing for publication in the most prestigious journals. This combination of culture and incentives inevitably leads to *all* analyses being portrayed as confirmatory and hypothesis driven even where many were exploratory. In this way, authors can generate a product that is attractive to journals while also maintaining the illusion (and possibly *de*lusion) that they have adhered to the hypothetico-deductive model of the scientific method.

The decision space in which these exploratory analyses reside is referred to as "researcher degrees of freedom." Beyond the exclusion of outliers, it can include decisions such as which conditions to enter into a wider analysis of multiple factors, which covariates or regressors to take into account, whether or not to collect additional participants, and even how to define the dependent measure itself. In even the simplest experimental design, these pathways quickly branch out to form a complex decision tree that a researcher can navigate either deliberately or unconsciously in order to generate statistically significant effects. By selecting the most desirable outcomes, it is possible to reject H0 in almost any set of data—and by combining selective reporting with HARKing (as described in chapter 1) it is possible to do so in favor of almost any alternative hypothesis.

Exploiting researcher degrees of freedom to generate statistical significance is known as "*p*-hacking" and was brought to prominence in 2011 by Joe Simmons, Leif Nelson, and Uri Simonsohn from the University of Pennsylvania and University of California Berkeley.[4] Through a combination of real experiments and simulations, Simmons and colleagues showed how selective reporting of exploratory analyses can generate meaningless *p* values. In one such demonstration, the authors simulated a simple experiment involving one independent variable (the intervention), two dependent vari-

ables (behaviors being measured), and a single covariate (gender of the participant). The simulation was configured so that there was no effect of the manipulation, that is, H0 was predetermined to be true. They then simulated the outcome of the experiment 15,000 times and asked how often at least one statistically significant effect was observed ($p<.05$). Given that H0 was true in this scenario, a nominal rate of 5 percent false positives can be assumed at $\alpha = .05$. The central question posed by Simmons and colleagues was what would happen if they embedded hidden flexibility within the analysis decisions. In particular, they tested the effect of analyzing either of the two dependent variables (reporting if a positive effect was obtained on either one), including gender as a covariate or not, increasing the number of participants after analyzing the results, and dropping one or more of the conditions. Allowing maximal combinatorial flexibility between these four options increased the false positive rate from a nominal 5 percent to an alarming 60.7 percent.

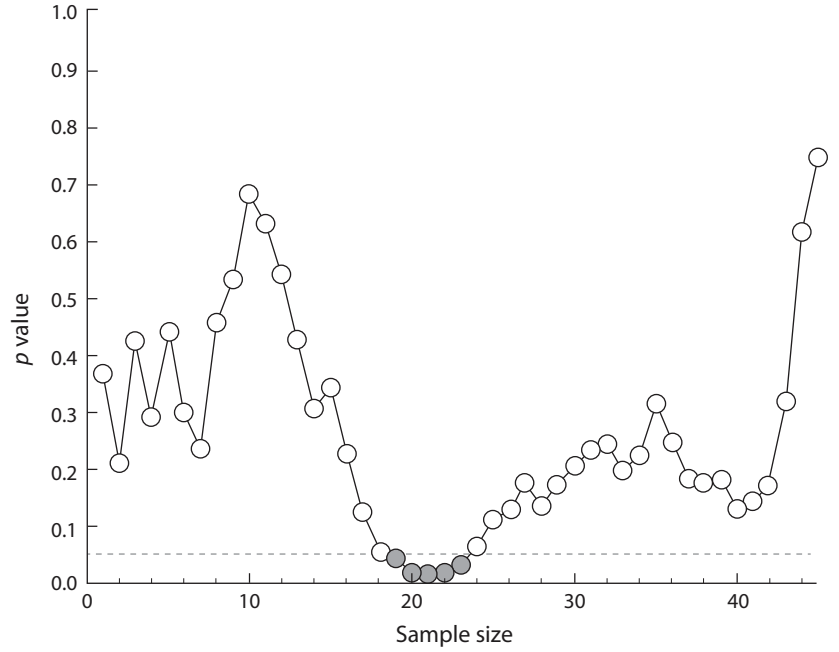
As striking as this is, 60.7 percent is probably still an underestimate of the true rate of false positives in many psychology experiments. Simmons and colleagues didn't even include other common forms of hidden flexibility, such as variable criteria for excluding outliers or conducting exploratory analyses within subgroups (e.g., males only or females only). Their simulated experimental design was also relatively simple and produced only a limited range of researcher degrees of freedom. In contrast, many designs in psychology are more complicated and will include many more options. Using a standard research design with four independent variables and one dependent variable, psychologist Dorothy Bishop has showed that at least one statistically significant main effect or interaction can be expected by chance in more than 50 percent of analyses—an order of magnitude higher than the conventional α threshold.[5] Crucially, this rate of false positives occurs even without exploiting the researcher degrees of freedom illustrated by Simmons and colleagues. Thus, where *p*-hacking occurs in more complex designs it is likely to render the obtained *p* values completely worthless.

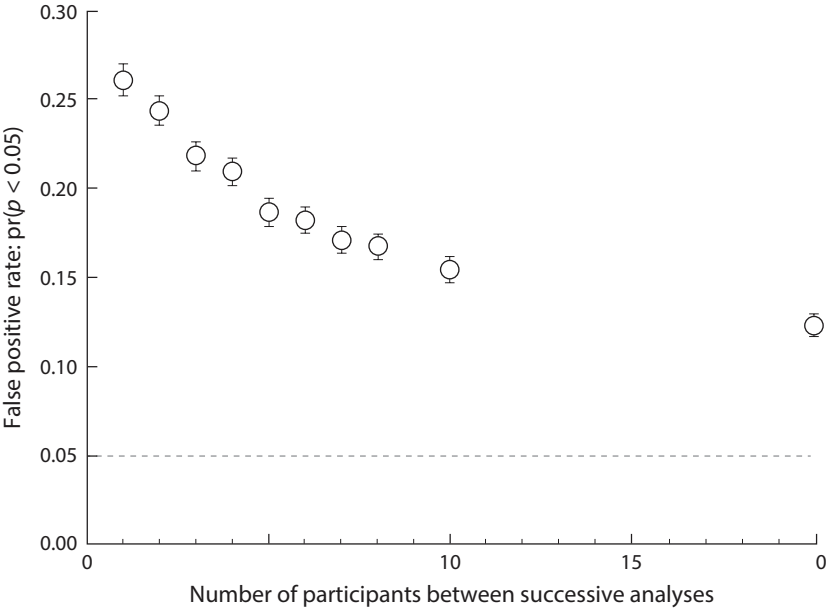
One key source of hidden flexibility in the simulations by Simmons and colleagues was the option to add participants after inspecting the results. There are all kinds of reasons why researchers peek at data before data collection is complete, but one central motivation is efficiency: in an environment with limited resources it can often seem sensible to stop data collec-

tion as soon as all-important statistical significance is either obtained or seems out of reach. This temptation to peek and chase $p<.05$ is of course motivated by the fact that psychology journals typically require the main conclusions of a paper to be underpinned by statistically significant results. If a critical statistical test returns $p = .07$, the researcher knows that reviewers and editors will regard the result as weak and unconvincing, and that the paper has little chance of being published in a competitive journal. Many researchers will therefore add participants in an attempt to nudge the *p* value over the line, without reporting in the published paper that they did so.

This kind of behavior may seem rational within a publishing system where what is best for science conflicts with the incentives that drive individual scientists. After all, if scientists are engaging in these practices then it is surely because they believe they have no other choice in the race for jobs, grant funding, and professional esteem. Unfortunately, however, chasing statistical significance by peeking and adding data completely undermines the philosophy of NHST. A central but often overlooked requirement of NHST is that researchers prespecify a stopping rule, which is the final sample size at which data collection must cease. Peeking at data prior to this end point in order to decide whether to continue or to stop increases the chances of a false positive. This is because NHST estimates the probability of the observed data (or more extreme) under the null hypothesis, and since the aggregate data pattern can vary randomly as individual data points are added, repeated hypothesis testing increases the odds that the data will, by chance alone, fall below the nominated α level (see figure 2.1). This increase in false positives is similar to that obtained when testing multiple hypotheses simultaneously.

Just how common is *p*-hacking? The lack of transparency in the research community makes a definitive answer impossible to glean, but important clues can be found by returning to the 2012 study by Leslie John and colleagues from chapter 1. Based on a survey of more than 2,000 American psychologists, they estimated that 100 percent have, on at least one occasion, selectively excluded data after looking at the impact of doing so, and that 100 percent have collected more data for an experiment after seeing whether results were statistically significant. They also estimate that 75 percent of psychologists have failed to report all conditions in an experiment, and that more than 50 percent have stopped data collection after





achieving a "desired result." These results indicate that, far from being a rare practice, *p*-hacking in psychology may be the norm.

Peculiar Patterns of *p*

The survey by John and colleagues suggests that *p*-hacking is common in psychology, but can we detect more objective evidence of its existence? One possible clue lies in the way *p* values are reported. If *p*-hacking is as common as claimed, then it should distort the distribution of *p* values in published work. To illustrate why, consider the following scenarios. In one, a team of researchers collect data by adding one participant at a time and successively analyze the results after each participant until statistical significance is obtained. Given this strategy, what *p* value would you expect to see at the end of the experiment? In another scenario, the team obtain a *p* value of .10 but don't have the option to collect additional data. Instead, they try ten different methods for excluding statistical outliers. Most of these produce *p* values higher than .05, but one reduces the *p* value to .049. They therefore select that option in the declared analysis and don't report the other "failed" attempts. Finally, consider a situation where $p = .08$ and the researchers have several degrees of freedom available, in terms of char-

FIGURE 2.1. The peril of adopting a flexible stopping rule in null hypothesis significance testing. In the **upper panel**, an experiment is simulated in which the null hypothesis (H0) is true and a statistical test is conducted after each new participant up to a maximum sample size of 50. A researcher who strategically *p*-hacks would stop as soon as *p* drops below .05 (dotted line). In this simulation, *p* crosses the significance threshold after collecting data for 19 participants (red symbols), despite the fact that there is no real effect to be discovered. In the **lower panel** we see how the frequency of interim analyses influences the false positive rate, defined here as the probability of finding a *p* value < .05 before reaching the maximum sample size. Each symbol in this plot is the average false positive rate across 10,000 simulated experiments, based on a variable stopping rule when H0 is true. If the researcher initially collects data for five participants and then reanalyses after every subsequent set of 20 participants, the false positive rate is 0.12 (rightmost symbol), which is roughly twice the α threshold of .05 (dotted line). Moving leftward on the plot, as successive analyses become more frequent (i.e., peeking after fewer and fewer participants), the false positive rate increases. If the researcher checks after every single participant then it climbs to as high as 0.26, yielding a 1 in 4 chance that a positive result ($p<.05$) would be falsely declared where none truly exists.

acterizing the dependent variable—specifically, they are able report the results either in terms of response times or performance accuracy, or an integrated measure of both. After analyzing all these different measures they find that the integrated measure "works best," revealing a *p* value of .037, whereas the individual measures alone reveal only nonsignificant effects ($p>.05$).

Although each of these scenarios is different, they all share one thing: in each case the researcher is attempting to push the *p* value *just* over the line. If this is your goal then it would make sense to stop *p*-hacking as soon as the *p* value drops below .05—after all, why spend additional resources only to risk that an effect that is currently publishable might "disappear" with the addition of more participants or by looking at the data in a different way? By focusing on merely crossing the significance threshold, the outcome should be to create a cluster of *p* values just below .05.

A number of individual cases of such behavior have been alleged. In a *Science* paper published in 2012, researchers presented evidence that religious beliefs could be reduced by instructing people to complete a series of tasks that require rational, analytic thinking. Despite sample sizes across four experiments ranging from 57 to 179, each experiment returned *p* values within a range of $p = .03$ to $p = .04$. Critics have argued either that the authors knew precisely how many participants would be required in each experiment to attain statistical significance or that they *p*-hacked, consciously or unconsciously, in order to always narrowly reach it.[6] There is, however, an entirely innocent explanation. Through no fault of the authors, their paper could be one of many unbiased studies considered by *Science*, with the journal selectively publishing the one that "struck gold" in finding a sequence of four statistically significant effects. Where it is impossible to distinguish biased practices by researchers from publication bias by journals, the authors naturally deserve the benefit of the doubt.

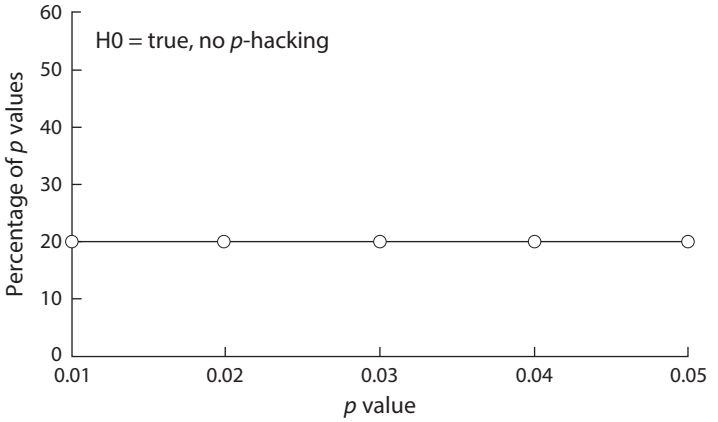
Because of the difficulty in differentiating publication bias from researcher bias, individual cases of *p*-hacking are difficult, if not impossible, to prove. However, if *p*-hacking is the norm then such cases may accrue across the literature to produce an overall preponderance of *p* values just below the significance threshold. In 2012, psychologists E. J. Masicampo and Daniel Lalande asked this question for the first time by examining the distribution of 3,627 *p* values sampled from three of the most prestigious psychology journals. Overall they found that smaller *p* values were more

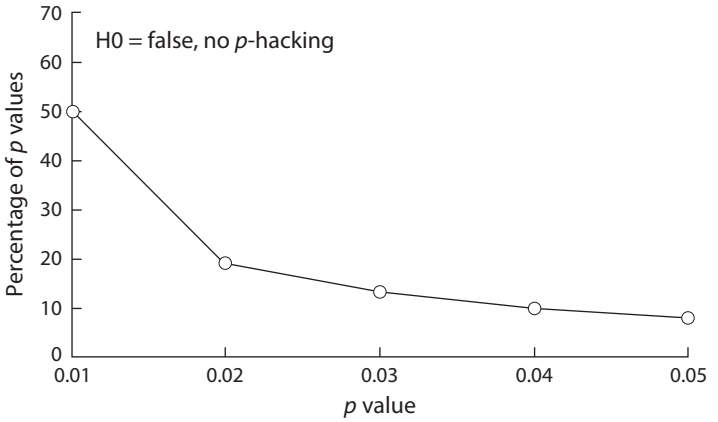
likely to be published than larger ones, but they also discovered that the number of *p* values just below .05 was about five times higher than expected.[7]

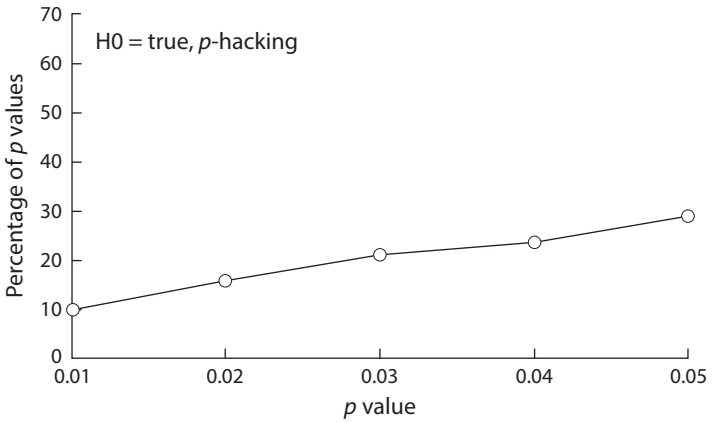
Masicampo and Lalande's findings have since been replicated by Nathan Leggett and colleagues from the University of Adelaide. Not only did they find the same spike in *p* values just below .05, but they also showed that the spike increased between 1965 and 2005.[8] The reason for growing numbers of "just-significant" results is not known for certain (and has itself been robustly challenged),[9] but if it is a genuine phenomenon then one possible explanation is the huge advancement in computing technology and statistical software. Undertaking NHST in 1965 was cumbersome and laborious (and often done by hand), which acted as a natural disincentive toward *p*-hacking. In contrast, modern software packages such as SPSS and R can reanalyze data many ways in just seconds.

It has been suggested that the studies of the Masicampo team and the Leggett team reveal evidence of *p*-hacking on a massive scale across thousands of studies, but is it possible to show such effects within more specific fields? A tool developed by Simonsohn, Leif, and Simmons called "*p*-curve" analysis promises to do just this.[10] The logic of *p*-curve is that the distribution of statistically significant *p* values within a set of studies reveals their evidential value (see figure 2.2). For unbiased (non-*p*-hacked) results where H0 is false, we should see more *p* values clustered toward the lower end of the spectrum (e.g., $p<.01$) than immediately below the significance threshold (e.g., *p* values between .04 to .05). This, in turn, should produce a distribution of *p* values between $p = 0$ and $p = .05$ that is positively skewed. In contrast, when researchers engage in *p*-hacking we should see a clustering of *p* values that is greatest just below .05, with fewer instances at lower values—thus the distribution of *p* values should be negatively skewed. Although *p*-curve has attracted controversy, it is a promising addition to the existing array of tools to detect hidden analytic flexibility.[11]

The problem of *p*-hacking is not unique to psychology. Compared with typical behavioral experiments, functional brain imaging includes far more researcher degrees of freedom. As the blogger Neuroskeptic has pointed out, the decision space of even the simplest fMRI study can include hundreds of analysis options, providing ample room for *p*-hacking.[12] At the time of writing, no studies of the distribution of *p* values have yet been undertaken for fMRI or electroencephalography (EEG), but indirect evidence







suggests that *p*-hacking may be just as common in these fields as in psychological science. Josh Carp of the University of Michigan has reported that out of 241 randomly selected fMRI studies, 207 employed unique analysis pipelines; this implies that fMRI researchers have numerous defensible options at their disposal and are making those analysis decisions after inspecting the data. As expected by a culture of *p*-hacking, earlier work has shown that the test-retest reliability of fMRI is moderate to low, with an estimated rate of false positives within the range of 10–40 percent.[13] We will return to problems with unreliability in chapter 3, but for now it is sufficient to note that *p*-hacking presents a serious risk to the validity of both psychology and cognitive neuroscience.

Institutionalized *p*-hacking damages the integrity of science and may be on the rise. If virtually all psychologists engage in *p*-hacking (even unconsciously) at least some of the time, and if *p*-hacking increases the rate of false positives to 50 percent or higher, then much of the psychological literature will be false to some degree. This situation is both harmful and, crucially, preventable. It prompts us to consider how the scientific record would change if *p*-hacking wasn't the norm and challenges us to reflect on how such practices can be tolerated in any community of scientists. Unfortunately, *p*-hacking appears to have crept up on the psychological community and become perceived as a necessary evil—the price we must pay to publish in the most competitive journals. Frustration with the practice is, however, growing. As Uri Simonsohn said in 2012 during a debate with fellow psychologist Norbert Schwarz:

> I don't know of anybody who runs a study, conducts one test, and publishes it no matter what the *p*-value is. . . . We are all *p* hackers, those of us who realize it want change.[14]

Figure 2.2. The logic of the *p*-curve tool developed by Uri Simonsohn and colleagues. Each plot shows a hypothetical distribution of *p* values between 0 and .05. For example, the *x*-value of .05 corresponds to all *p* values between .04 and .05, while .01 corresponded to all *p* values between 0 and .01. In the upper panel, the null hypothesis (H0) is true, and there is no *p*-hacking; therefore *p* values in this plot are uniformly distributed. In the middle panel, H0 is false, leading to a greater number of smaller *p* values than larger ones. The positive (rightward) skew in this plot doesn't rule out the presence of *p*-hacking but does suggest that the sample of *p* values is likely to contain evidential value. In the lower panel, H0 is true and more *p* values are observed closer to .05. The negative (leftward) skew in this plot suggests the presence of *p*-hacking.

Ghost Hunting

Could a greater emphasis on replication help solve the problems created by *p*-hacking?[15] In particular, if we take a case where a *p*-hacked finding is later replicated, does the fact that the replication succeeded mean we don't need to worry whether the original finding exploited researcher degrees of freedom?

If we assume that true discoveries will replicate more often than false discoveries then a coordinated program of replication certainly has the potential to weed out *p*-hacked findings, provided that the procedures and analyses of the original study and the replication are identical. However, the argument that replication neutralizes *p*-hacking has two major shortcomings. The first is that while direct (close) replication is vital, even a widespread and systematic replication initiative wouldn't solve the problem that *p*-hacked studies already waste resources by creating blind alleys. Second, as we saw earlier and will see again later, such a direct replication program simply doesn't exist in psychology. Instead, the psychological community has come to rely on the more loosely defined "conceptual replication" to validate previous findings, which satisfies the need for journals to publish novel and original results. Within a system that depends on conceptual replication, researcher degrees of freedom can be just as easily exploited to "replicate" a false discovery as to create one from scratch. A *p*-hacked conceptual replication of a *p*-hacked study tells us very little about reality apart from our ability to deceive ourselves.

The case of John Bargh and the elderly priming effect provides an interesting case where researcher degrees of freedom may have led to so-called phantom replication. Recall that at least two attempts to exactly replicate the original elderly priming effect have failed.[16] By way of rebuttal, Bargh has argued that two other studies did successfully replicate the effect.[17] However, if we look closely at those studies, we find that in neither case was there an overall effect of elderly priming—in one study the effect was statistically significant only once participants were divided into subgroups of low or high self-consciousness; and in the other study the effect was only significant when dividing participants into those with positive or negative attitudes toward elderly people. Furthermore, each of these replications used different methods for handling statistical outliers, and each analysis included covariates that were not part of the original elderly priming experi-

ments. These differences hint at a potential phantom replication. Driven by the confirmation bias to replicate the original (high-profile) elderly priming effect, the researchers in these subsequent studies may have consciously or unconsciously exploited researcher degrees of freedom to produce a successful replication and, in turn, a more easily marketable publication.

Whether these conceptual replications were contaminated by *p*-hacking cannot be known for certain, but we have good reason to be suspicious. Despite one-time prevalence estimates approaching 100 percent, researchers usually deny that they *p*-hack.[18] In Leslie John's survey in 2012, only about 60 percent of psychologists admitted to "Collecting more data after seeing whether results were significant," whereas the prevalence estimate derived from this admission estimate was 100 percent. Similarly, while ~30 percent admitted to "Failing to report all conditions" and ~40 percent admitted to "Excluding data after the impact of doing so," the estimated prevalence rates in each case were ~70 percent and ~100 percent, respectively. These figures needn't imply dishonesty. Researchers may sincerely deny *p*-hacking yet still do it unconsciously by failing to remember and document all the analysis decisions made after inspecting data. Some psychologists may even do it consciously but believe that such practices are acceptable in the interests of data exploration and narrative exposition. Yet, regardless of whether researchers are *p*-hacking consciously or unconsciously, the solution is the same. The only way to verify that studies are not *p*-hacked is to show that the authors planned their methods and analysis before they analyzed the data—and the only way to prove that is through study preregistration.

Unconscious Analytic "Tuning"

What do we mean exactly when we say *p*-hacking can happen unconsciously? Statisticians Andrew Gelman and Eric Loken have suggested that subtle forms of *p*-hacking and HARKing can join forces to produce false discoveries.[19] In many cases, they argue, researchers may behave completely honestly, believing they are following best practice while still exploiting researcher degrees of freedom.

To illustrate how, consider a scenario where a team of researchers design an experiment to test the a priori hypothesis that listening to classical music

improves attention span. After consulting the literature they decide that a *visual search* task provides an ideal way of measuring attention. The researchers choose a version of this task in which participants view a screen of objects and must search for a specific target object among distractors, such as the letter "O" among many "Q"s. On each trial of the task, the participant judges as quickly as possible whether the "O" is present or absent by pressing a button—half the time the "O" is present and half the time it is absent. To vary the need for attention, the researchers also manipulate the number of distractors (Qs) between three conditions: 4 distractors (low difficulty, i.e., the "O" pops out when it is present), 8 distractors (medium difficulty) and 16 distractors (high difficulty). The key dependent variables are the reaction times and error rates in judging whether the letter "O" is present or absent. Most studies report reaction times as the measure for this task and find that reaction times increase with the number of distractors. Many studies also report error rates. The researchers decide to adopt a repeated measures design in which each participant performs this task twice, once while listening to classical music and once while listening to nonclassical music (their control condition). They decide to test 20 participants.

So far the researchers feel they have done everything right: they have a prespecified hypothesis, a task selected with a clear rationale, and a sample size that is on parity with previous studies on visual search. Once the study is complete, the first signs of their analysis are encouraging: they successfully replicate two effects that are typically observed in visual search tasks, namely that participants are significantly slower and more error prone under conditions with more distractors (16 is more difficult than 8 which, in turn, is more difficult than 4), and that they are significantly slower to judge when a target is absent compared to when it is present. So far so good. On this basis, the authors judge that the task successfully measured attention.

But then it gets trickier. The researchers find no statistically significant main effect of classical music on either reaction times or error rates, which does not allow them to reject the null hypothesis. However they do find a significant interaction between the type of music (classical, nonclassical) and the number of distractors (4, 8, 16) for error rates ($p = .01$) but not for reaction times ($p = .7$). What this interaction means is that, for error rates,

the effect of classical music differed significantly between the different distractor conditions. Post hoc comparisons show that error rates were significantly reduced when participants were exposed to classical music compared with control music, in displays with 16 distractors ($p = .01$) but not in displays with 8 or 4 distractors (both $p>.2$). In a separate analysis they also find that reaction times on target absent trials (i.e., no "O" present) are significantly faster when exposed to classical music compared with control music ($p = .03$). The same advantage isn't significant for trials where the "O" was present ($p = .55$) or when target-present and target-absent trials are averaged together (as shown by the lack of a significant main effect).

The researchers think carefully about their results. After doing some additional reading they learn that error rates can sometimes provide more sensitive measures of attention than reaction times, which would explain why classical music influenced only error rates. They also know that the "target absent" condition is more difficult for participants and therefore perhaps a more sensitive measure of attentional capacity—that would also explain why classical music boosted performance on trials without targets. Finally, they are pleased to see that the benefit of classical music on error rates with 16 distractors goes in the direction predicted by their hypothesis. Therefore, despite the fact that the main effect of classical music is not statistically significant on either of the measures (reaction times or error rates), the researchers conclude that the results support the hypothesis: classical music improves visual attention, *particularly under conditions of heightened task difficulty*. In the introduction of their article they phrase their hypothesis as, "We predicted that classical music would improve visual search performance. Since misidentification of visual stimuli can provide a more sensitive measure of attention than reaction time (e.g., Smith 2000), such effects may be expected to occur most clearly in error rates, especially under conditions of heightened attentional load or task difficulty." In the discussion section of the paper, the authors note that their hypothesis was supported and argue that their results conceptually replicate a previous study, which showed that classical music can improve the ability to detect typographic errors in printed text.

What, if anything, did the researchers do wrong in this scenario? Many psychologists would argue that their behavior is impeccable. After all, they didn't engage in questionable practices such as adding participants until

statistical significance was obtained, selectively removing outliers, or testing the effect of various covariates on statistical significance. Moreover, they had an a priori hypothesis, which they tested, and they employed a manipulation check to confirm that their visual search task measured attention. However, as Gelman and Loken point out, the situation isn't so simple—researcher degrees of freedom have still crept insidiously into their conclusions.

The first problem is the lack of precision in the researchers' a priori hypothesis, which doesn't specify the dependent variable that should show the effect of classical music (either reaction time or error rates, or both) and doesn't state under what conditions that hypothesis would or would not be supported. By proposing such a vague hypothesis, the researchers have allowed any one of many different outcomes to support their expectations, ignoring the fact that doing so inflates the Type I error rate (α) and invites confirmation bias. Put differently, although they have an a priori scientific hypothesis, it is consistent with multiple *statistical* hypotheses.

The second problem is that the researchers ignore the lack of a statistically significant main effect of the intervention on either of the measures; instead they find a significant interaction between the main manipulation (classical music vs. control music) and the number of distractors (4, 8, 16)—but for error rates only. Since the researchers had no specific hypothesis that the effect of classical music would increase at greater distractor set sizes *for error rates only*, this result was unexpected. Yet by framing this analysis as though it was a priori, the researchers tested more null hypotheses than were implied in their original (vague) hypothesis, which in turn increases the chances of a false positive.

Third, the researchers engage in HARKing. Even though their general hypothesis was decided before conducting the study, it was then refined based on the data and is presented in the introduction in its refined state. This behavior is a subtle form of HARKing that conflates hypothesis testing with post hoc explanation of unexpected results. Since the researchers did have an a priori hypothesis (albeit vague) they would no doubt deny that they engaged in HARKing. Yet even if blurred in their own recollections, the fact is that they adjusted and refined their hypothesis to appear consistent with unexpected results.

Finally, despite the fact that their findings are less definitive than advertised, the researchers treat them as a conceptual replication of previous

work—a corroboration of the general idea that exposure to classical music improves attention. Interestingly, it is in this final stage of the process where the lack of precision in their original hypothesis is most clearly apparent. This practice also highlights the inherent weakness of conceptual replication, which risks constructing bodies of knowledge on weak units of evidence.

Is this scenario dishonest? No. Is it fraudulent? No. But does it reflect questionable research practices? Yes. Unconscious as it may be, the fact is that the researchers in this scenario allowed imprecision and confirmation bias to distort the scientific record. Even among honest scientists, researcher degrees of freedom pose a serious threat to discovery.

Biased Debugging

Sometimes hidden flexibility can be so enmeshed with confirmation bias that it becomes virtually invisible. In 2013, Mark Stokes from the University of Oxford highlighted a situation where an analysis strategy that seems completely sensible can lead to publication of false discoveries.[20] Suppose a researcher completes two experiments. Each experiment involves a different method to provide convergent tests for an overarching theory. In each case the data analysis is complicated and requires the researcher to write an automated script. After checking through each of the two scripts for obvious mistakes, the researcher runs the analyses. In one of the experiments the data support the hypothesis indicated by the theory. In the second experiment, however, the results are inconsistent. Puzzled, the researcher studies the analysis script for the second experiment and discovers a subtle but serious error. Upon correcting the error, the results of the second experiment become consistent with the first experiment. The author is delighted and concludes that the results of both experiments provide convergent support for the theory in question.

What's wrong with this scenario? Shouldn't we applaud the researcher for being careful? Yes and no. On the one hand, the researcher has caught a genuine error and prevented publication of a false conclusion. But note that the researcher didn't bother to double-check the analysis of the *first* experiment because the results in that case turned out as expected and—perhaps more importantly—as desired. Only the second experiment at-

tracted scrutiny because it ran counter to expectations, and running counter to expectations was considered sufficient grounds to believe it was erroneous. Stokes argues that this kind of results-led debugging—also termed "selective scrutiny"—threatens to magnify false discoveries substantially, especially if it occurs across an entire field of research.[21] And since researchers never report which parts of their code were debugged or not (and rarely publish the code itself), biased debugging represents a particularly insidious form of hidden analytic flexibility.

Are Research Psychologists Just Poorly Paid Lawyers?

The specter of bias and hidden analytic flexibility inevitably prompts us to ask: what is the job of a scientist? Is it to accumulate evidence as dispassionately as possible and decide on the weight of that evidence what conclusion to draw? Or is it our responsibility to advocate a particular viewpoint, seeking out evidence to support that view? One is the job of a scientist; the other is the job of a lawyer. As psychologist John Johnson from Pennsylvania State University said in a 2013 blog post at *Psychology Today*:

> Scientists are not supposed to begin with the goal of convincing others that a particular idea is true and then assemble as much evidence as possible in favor of that idea. Scientists are supposed to be more like detectives looking for clues to get to the bottom of what is actually going on. They are supposed to be willing to follow the trail of clues, wherever that may lead them. They are supposed to be interested in searching for the critical data that will help decide what is actually true, not just for data that supports a preconceived idea. Science is supposed to be more like detective work than lawyering.[22]

Unfortunately, as we have seen so far, psychology falls short of meeting this standard. Whether conscious or unconscious, the psychological community tortures data until the numbers say what we want them to say—indeed what many psychologists, deep down, would admit we *need* them to say in the competition for high-impact publications, jobs, and grant funding. This situation exposes a widening gulf between the needs of the scientists and the needs of science. Until these needs are aligned in favor of science and the public who fund it, the needs of scientists will always win—to our own detriment and to that of future generations.

Solutions to Hidden Flexibility

Hidden flexibility is a problem for any science where the act of discovery involves the accumulation of evidence. This process is less certain in some sciences than in others. If your evidence is clearly one way or the other, such as the discovery of a new fossil or galaxy, then inferences from statistics may be unnecessary. In such cases no analytic flexibility is required, whether hidden or disclosed—so none takes place. It is perhaps this association between statistical analysis and noisy evidence that prompted physicist Ernest Rutherford to allegedly once remark: "If your experiment needs statistics, you ought to have done a better experiment."

In many life sciences, including psychology, discovery isn't a black-and-white issue; it is matter of determining, from one experiment to the next, the theoretical contribution made by various shades of gray. When psychologists set arbitrary criteria ($p<.05$) on the precise shade of gray required to achieve publication—and hence career success—they also incentivize a host of conscious and unconscious strategies to cross that threshold. In the battle between science and storytelling, there is simply no competition: storytelling wins every time.

How can we get out of this mess? Chapter 8 will outline a manifesto for reform, many aspects of which are already being adopted. The remainder of this chapter will summarize some of the methods we can use to counter hidden flexibility.

Preregistration. The most thorough solution to *p*-hacking and other forms of hidden flexibility (including HARKing) is to prespecify our hypotheses and primary analysis strategies before examining data. Preregistration ensures that readers can distinguish the strategies that were independent of the data from those that were (or might have been) data led. This is not to suggest that data-led strategies are necessarily incorrect or misleading. Some of the most remarkable advances in science have emerged from exploration, and there is nothing inherently wrong with analytic flexibility. The problems arise when that flexibility is hidden from the reader, and possibly from the researcher's own awareness. By unmasking this process, preregistration protects the outcome of science from our own biases as human practitioners.

Study preregistration has now been standard practice in clinical medicine for decades, driven by concerns over the effects of hidden flexibility

and publication bias on public health. For basic science, these risks may be less immediate but they are no less serious. Hidden flexibility distorts the scientific record, and, since basic research influences and feeds into more applied areas (including clinical science), corruption of basic literature necessarily threatens any forward applications of discovery.

Recent years have witnessed a concerted push to highlight the benefits of preregistration in psychology. An increasing number of journals are now offering preregistered article formats in which part of the peer review process happens before researchers conduct experiments. This type of review ensures adherence to the hypothetico-deductive model of the scientific method, and it also prevents publication bias. Resources such as the Open Science Framework also provide the means for researchers to preregister their study protocols.

***p*-curve.** The *p*-curve tool developed by Simonsohn and colleagues is useful for estimating the prevalence of *p*-hacking in published literature. It achieves this by assuming that collections of studies dominated by *p*-hacking will exhibit a concentration of *p* values that peaks just below .05. In contrast, an evidence base containing positive results in which *p*-hacking is rare or absent will produce a positively skewed distribution where smaller *p* values are more common than larger ones. Although this tool cannot diagnose *p*-hacking within individual studies, it can tell us which fields within psychology suffer more from hidden flexibility. Having identified them, the community can then take appropriate corrective action such as an intensive program of direct replication.

Disclosure statements. In 2012, Joe Simmons and colleagues proposed a straightforward way to combat hidden flexibility: simply ask the researchers.[23] This approach assumes that most researchers (a) are inherently honest and will not deliberately lie, and (b) recognize that exploiting researcher degrees of freedom reduces the credibility of their own research. Therefore, requiring researchers to state whether or not they engaged in questionable practices should act as a disincentive to *p*-hacking, except for researchers who are either willing to admit to doing it (potentially suffering loss of academic reputation) or are willing to lie (active fraud).

The disclosure statements suggested by the Simmons team would require authors to state in the methods section of submitted manuscripts how they

made certain decisions about the study design and analysis. These include whether the sample size was determined before the study began and if any of experimental conditions or data were excluded. Their "21 word solution" (as they call it) is:

> We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Although elegant in their simplicity, disclosure statements have several limitations. They can't catch forms of *p*-hacking that exploit defensible ambiguities in analysis decisions, such as testing the effect of adding a covariate to the design or focusing the analysis on a particular subgroup (e.g., males only) following exploration. Furthermore, the greater transparency of methods, while laudable in its own right, doesn't stop the practice of HARKing; note that researchers are not asked whether their a priori hypotheses were altered after inspecting the data. Finally, disclosure statements cannot catch unconscious exploitation of researcher degrees of freedom, such as forgetting the full range of analyses undertaken or more subtle forms of HARKing (as proposed by Gelman and Loken). Notwithstanding these limitations, disclosure statements are a worthy addition to the range of tools for counteracting hidden analytic flexibility.

Data sharing. The general lack of data transparency is a major concern in psychology, and will be discussed in detail in chapter 4. For now, it is sufficient to note that data sharing provides a natural antidote to some forms of *p*-hacking—particularly those where statistical significance is based on post hoc analytic decisions such as different methods for excluding outliers or focusing on specific subgroups. Publishing the raw data allows independent scientists with no vested interest to test how robust the outcome is to alternative analysis pathways. If such examinations were to reveal that the authors' published approach was the only one out of a much larger subset to produce $p<.05$, the community would be justifiably skeptical of the study's conclusions. Even though relatively few scientists scrutinize each other's raw data, the mere possibility that this could happen should act as a natural deterrent to deliberate *p*-hacking.

Solutions to allow "optional stopping." Leslie John's survey showed how a major source of *p*-hacking is violation of stopping rules; that is, continu-

ously adding participants to an experiment until the *p* value drops below the significance threshold. When H0 is true, *p* values between 0 and 1 are all equally likely to occur, therefore with the addition of enough participants a *p* value below .05 will eventually be found by chance. Psychologists often neglect stopping rules because, in most studies, there is no strong motivation for selecting a particular sample size in the first place.

Fixed stopping rules present a problem for psychology because the size of the effect being investigated is often small and poorly defined. Fortunately, there are two solutions that psychologists can use to avoid violating stopping rules. The first, highlighted by Michael Strube from Washington University (and more recently by Daniël Lakens), allows researchers to use a variable stopping rule with NHST by lowering the α level in accordance with how regularly the researcher peeks at the results.[24] This correction is similar to more conventional corrections for multiple comparisons. The second approach is to adopt Bayesian hypothesis testing in place of NHST.[25] Unlike NHST, which estimates the probability of observed (or more extreme) data under the null hypothesis, Bayesian tests estimate the relative probabilities of the data under competing hypotheses. In chapter 3 we will see how this approach has numerous advantages over more conventional NHST methods, including the ability to directly estimate the probability of H0 relative to H1. Moreover, Bayesian tests allow researchers to sequentially add participants to an experiment until the weight of evidence supports either H0 or H1. According to Bayesian statistical philosophy "there is nothing wrong with gathering more data, examining these data, and then deciding whether or not to stop collecting new data—no special corrections are needed."[26] This flexibility is afforded by the likelihood principle, which holds that "[w]ithin the framework of a statistical model, all of the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses."[27] In other words, the more data you have, the more accurately you can conclude support for one hypothesis over another.

Standardization of research practices. One reason *p*-hacking is so common is that arbitrary decisions are easy to justify. Even within a simple experiment, there are dozens of different analytic pathways that a researcher can take, all of which may have a precedent in the published literature and all of which may be considered defensible. This ambiguity

enables researchers to pick and choose the most desirable outcome from a smorgasbord of statistical tests, reporting the one that "worked" as though it was the only analysis that was attempted. One solution to this problem is to apply more constraints on the range of acceptable approaches, applying institutional standards to such practices as outlier exclusion or the use of covariates.

Scientists are generally resistant to such changes, particularly when a clear best practice fails to stand out from the crowd. Such standards can also be difficult to apply in emerging fields, such as functional brain imaging, owing to the rapid developments in methodology and analysis strategies. However, where standardization isn't possible and any one of several arbitrary decisions is possible, there is a strong argument that researchers should report all of them and then summarize the robustness of the outcomes across all contingencies.

Moving beyond the moral argument. Is *p*-hacking a form of fraud? Whether questionable research practices such as *p*-hacking and HARKing are fraudulent—or even on the same continuum as fraud—is controversial. Psychologist Dave Nussbaum from the University of Chicago has argued that clearly fraudulent behavior, such as data fabrication, is categorically different from questionable practices such as *p*-hacking because, with *p*-hacking, the intent of the researcher cannot be known.[28] Nussbaum is right. As we have seen in this chapter, many cases of *p*-hacking and HARKing are likely to be unconscious acts of self-deception. Yet in cases where researchers deliberately *p*-hack in order to achieve statistical significance, Nussbaum agrees that such behavior is on the same continuum of extreme fraud.

We will return to the issue of fraud, but for now we can ask: does it matter if *p*-hacking is fraudulent? Whether the sin of hidden flexibility is deliberate or an act of self-deception is a distraction from the goal of reforming science. Regardless of what lies in the minds of researchers, the effects of *p*-hacking are clear, populating the literature with post hoc hypotheses, false discoveries, and blind alleys. The solutions are also the same.

CHAPTER 3

The Sin of Unreliability

And it's this type of integrity, this kind of care not to fool yourself, that is missing to a large extent in much of the research in cargo cult science.

—Richard Feynman, 1974

Particles break light-speed limit," announced *Nature News*.[1] "Faster than light particles threaten Einstein," declared *Reuters*.[2] "Was Einstein wrong?" asked *Time*.[3] So read the headlines in September 2011 when a team of physicists published evidence suggesting that subatomic particles called neutrinos could travel faster than the speed of light. If true, this discovery would revolutionize modern physics. Teams of scientists immediately began the task of repeating the experiment. By June 2012, three independent groups had failed to replicate the original result: the neutrinos in their experiments traveled at approximately the speed of light, just as predicted by special relativity. One month later, the original team reported that their findings were caused by a loose fiber-optic cable.

The case of faster-than-light neutrinos may sound like an example of science going awry, but, in fact, it is just the opposite. After one team of scientists made what looked like an extraordinary discovery, the scientific community responded by attempting to reproduce the result. When those attempts failed, the original team scrutinized their experiment more closely, discovering a technical error that explained the anomaly. Science can never escape the risk of human error, but it can and must ensure that it self-corrects. Imagine for a moment what kind of physics we would have if faster-than-light neutrinos had simply been believed, their status as a discovery left unreplicated and unchallenged.

Replication is the immune system of science, identifying false discoveries by testing whether other scientists can repeat them. Without replication, we have no way of knowing which discoveries are genuine and which are caused by technical error, researcher bias, fraud, or the play of chance. And if we don't know which results are reliable, how can we generate meaningful theories?

Unfortunately, as we saw earlier, the process of replication—so intrinsic to the scientific method—is largely ignored or distorted in psychology. Recall from chapter 1 that the claims of psychic precognition by Daryl Bem soared into print at the one of the most prestigious psychology journals in the world. Yet the crucial nonreplication by Stuart Richie and Chris French took much longer to appear and was initially rejected, on principle alone, by the same journal that published Bem's findings. Instead of valuing the reproducibility of results, psychology has embraced a tabloid culture where

novelty and interest-value are paramount and the truth is left begging. Psychology thus succumbs to our third major transgression: the sin of unreliability.

Sources of Unreliability in Psychology

Science generates an understanding of the natural world by using empirical evidence to reduce uncertainty. The hypothetico-deductive model of the scientific method, introduced in chapter 1, achieves this by systematically testing hypotheses borne from theory. Once the results of high-quality experiments are verified through direct (close) replication, the evidence can refine the theory in question, generating further hypotheses and honing our ability to understand and predict reality (see figure 3.1).

Unfortunately, psychology fails to adhere to this philosophy, and nowhere is this culture expressed more blatantly than by the indifference—and, in many cases, hostility—toward direct replication. Yet the lack of replication isn't the only reason psychological science faces a crisis of reliability. As we will see, other concerns include low statistical power, failure to disclose full methodological details, statistical fallacies, and the refusal to retract irreproducible findings from the literature. Together these problems not only threaten the truth-value of psychological evidence; they threaten the status of psychology as a science.

Reason 1: Disregard for Direct Replication

Direct replication is intrinsic to all sciences. During a direct replication, a researcher seeks to the test the repeatability of a previous finding by duplicating the methodology as exactly as possible.[4] The importance of direct replication in science is reflected by the simple fact that empirical journal articles in all sciences include method sections. A method is supposed to provide researchers with all the information they would need to replicate the experimental "recipe."

Despite the clear importance of replication, we saw in chapter 1 how the academic culture in psychology places little emphasis in repeating the experimental methods of other psychologists. Such work is seen to lack in-

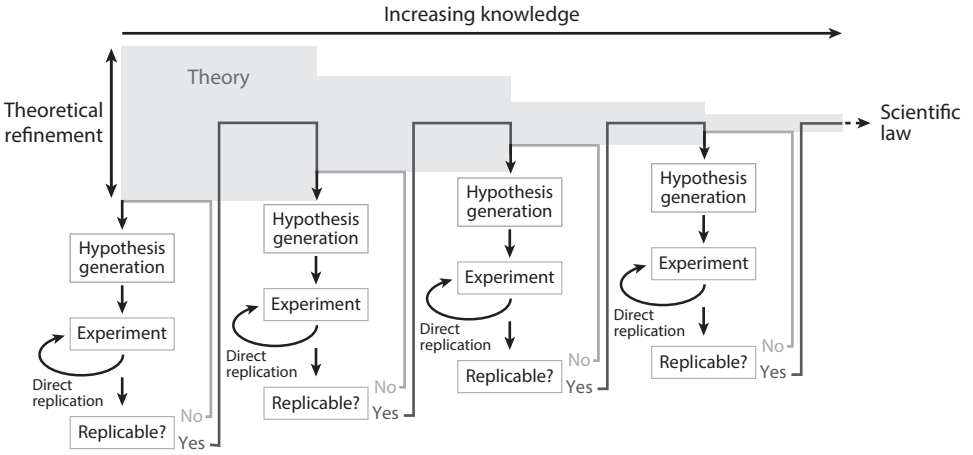


FIGURE 3.1. Science usually advances in steady increments rather than leaps. According to the deductive scientific method, a hypothesis is generated from current theory, and an experiment is designed to test the hypothesis against one or more competitors. If the outcome of the experiment can be directly replicated to the community's satisfaction, then the theory is refined, and a new hypothesis is formulated. The endpoint of this process of knowledge accumulation is the generation of scientific laws. Note that the refining of theoretical precision in no way implies that the original theoretical model remains intact or must be in any way sustained: the increase of knowledge can, and does, lead to theories being discarded altogether. Regardless, the theoretical framework always becomes more precise with the addition of new evidence.

novation within a system that instead seeks to validate previous findings by conducting novel experiments that test a related (but different) idea using a different method: the approach referred to as "conceptual replication." Although used widely in psychology, the term conceptual replication does not feature in the scientific method of other disciplines. In fact, the term itself is misleading because conceptual replications don't actually replicate previous experiments; they instead assume (rather than test) the truth of the findings revealed in those experiments, infer the underlying cause, and then seek converging evidence for that inferred cause using an entirely different experimental procedure. Viewed within the framework of the H-D scientific method, this process can be thought of as extrapolating from a body of findings to refine theory and generate new hypotheses (see figure 3.1). As important as this step is, it depends first and foremost on the reliability of the underlying evidence base. To rely solely on extrapolation at the expense of direct replication is to build a house on sand.

If direct replications are considered trivial and uninteresting then they should be seldom seen in the published literature. In his famous 1974 lecture "Cargo Cult Science," Richard Feynman recounts his experience of suggesting to a psychology student that she perform a direct replication of a previous experiment before attempting a novel one:

> She was very delighted with this new idea, and went to her professor. And his reply was, no, you cannot do that, because the experiment has already been done and you would be wasting time. This was in about 1947 or so, and it seems to have been the general policy then to not try to repeat psychological experiments, but only to change the conditions and see what happened.[5]

Remarkably, since 1947 there has been little analysis of how frequently psychologists replicate each other's work even partially, let alone directly. In 2012, Matthew Makel, Jonathan Plucker, and Boyd Hegarty conducted the first systematic investigation of replication rates in psychology.[6] They searched the top 100 psychology journals from 1900 to 2012 and found that only 1.57 percent of 321,411 articles mentioned a word beginning with "replicat*." This may sound low enough but was already optimistically high. Within a randomly selected subsample of 500 of the 1.57 percent, only 342 actually reported some form of replication—and, of these, just 62 articles reported a direct replication of a previous experiment. On top of that, only 47 percent of replications within the subsample were produced by independent researchers. The implications of the Makel study are sobering: for every 1,000 papers published in psychology, only two will seek to directly replicate a previous experiment, and just one of those will be published by a different team to the original study.

What happens to a scientific discipline when it abandons direct replication, as psychology has done? The immediate consequence, estimated by medical researcher John Ioannidis, is that up to 98 percent of published findings may be either unconfirmed genuine discoveries or unchallenged fallacies.[7] Failure to attempt replications thus sabotages the ability for psychological science to self-correct: where physicists succeeded (quickly) at refuting and explaining the observation of faster-than-light neutrinos, psychologists would fail by not trying in the first place. This contempt for verification, in turn, has a profound effect on theory development, condemning the field to the publication of theoretical frameworks that can be neither

confirmed nor falsified. In the words of psychologists Chris Ferguson and Moritz Heene, the outcome is a discipline that plays host to a "graveyard of undead theories," each of them as immovable as they are impotent.[8]

Why is psychology so implacably opposed to direct replication? The short answer is that we don't know for certain—the aversion is so entrenched that the causes and effects have become obscured. The longer answer, as discussed in chapter 1, is that together with other life sciences, psychology has evolved an incentive structure that rewards empiricists who can produce novel, positive, eye-catching results that confirm the hypothesis and offer pithy interpretations. Pressure *not* to replicate is applied from all directions. At the supply end, funding agencies are loath to award money for merely repeating previous research—in the UK, even grant applications that describe original research (let alone replications) are often rejected for lack of novelty and innovation. Meanwhile, at the demand end, it is a struggle to publish direct replications in respected journals, whether successful or not. When such attempts succeed they are generally seen as boring and contributing little of value ("we already knew this."; "what does this add?"), despite the fact that replications frame the certainty we can justify in prior discoveries. And when replications fail, defenders of the original work are liable to try and block publication or respond aggressively when such work is published.

The negative attitude toward replication in psychology is epitomized by an incident that became known rather infamously as "Repligate." In May 2014, the journal *Social Psychology* bucked the academic trend and reported an ambitious initiative to reproduce a series of influential psychological discoveries claimed since the 1950s.[9] Many of the findings could not be replicated, and in most cases these nonreplications were met with cordial interactions between researchers. For instance, confronted with the news that one of his prior findings in social priming could not be reproduced, Dr. Eugene Caruso from the University of Chicago said, "This was certainly disappointing at a personal level. But when I take a broader perspective, it's apparent that we can always learn something from a carefully designed and executed study."

Not all researchers were so gracious. Dr. Simone Schnall from the University of Cambridge argued that her work on social priming was treated unfairly and "defamed." In a remarkable public statement, Schnall claimed that she was bullied by the researchers who sought (unsuccessfully) to rep-

licate her findings and that the journal editors who agreed to publish the failed replications of her work behaved unethically.[10] She wrote, "I feel like a criminal suspect who has no right to a defence and there is no way to win: The accusations that come with a "failed" replication can do great damage to my reputation, but if I challenge the findings I come across as a 'sore loser.' "

Schnall's strong reaction to the failed replication of her own work provoked a mixed reaction from the psychological community. While many psychologists were bewildered by her response, a number of prominent US psychologists voiced support for her position. Dan Gilbert from Harvard University likened Schnall's battle to the plight of Rosa Parks,[11] and he referred to some psychologists who conducted or supported replications as "bullies," "replication police," "second stringers," McCarthyists, and "god's chosen soldiers in a great jihad."[12] Others accused the so-called replicators of being "Nazis," "fascists," and "mafia." Rather than viewing replication as an intrinsic part of best scientific practice, Gilbert and his supporters framed it as a threat to the reputation of the (presumably brilliant) researchers who publish irreproducible findings, stifling their creativity and innovation.[13]

For some psychologists, the reputational damage in such cases is grave—so grave that they believe we should limit the freedom of researchers to pursue replications. In the wake of Repligate, Nobel laureate Daniel Kahneman called for a new rule in which replication attempts should be "prohibited" unless the researchers conducting the replication consult beforehand with the authors of the original work.[14] Kahneman said, "Authors, whose work and reputation are at stake, should have the right to participate as advisers in the replication of their research." Why? Because the method sections published by psychology journals are too vague to provide a recipe that can be repeated by others. Kahneman argued that successfully reproducing original effects could depend on seemingly irrelevant factors—hidden secrets that only the original authors would know. "For example, experimental instructions are commonly paraphrased in the methods section, although their wording and even the font in which they are printed are known to be significant."

For many psychologists, Kahneman's cure is worse than the disease. Andrew Wilson from Leeds Metropolitan University immediately rejected

the suggestion of new regulations, writing: "If you can't stand the replication heat, get out of the empirical kitchen because publishing your work means you think it's ready for prime time, and if other people can't make it work based on your published methods then that's your problem and not theirs."[15] Are replications in other sciences ever regarded as an act of aggression, as Schnall and others suggest? Jon Butterworth, head of the Department of Physics and Astronomy at University College London, finds this view of replication completely alien. "Thinking someone's result is interesting and important enough to be checked is more like flattery," he told me. For Butterworth there is no question that the published methods of a scientific paper should be sufficient for trained specialists in the field to repeat experiments, without the need to learn unpublished secrets from the original authors. "Certainly no physicist I know would dare claim their result depended on hidden 'craft.' "[16]

Helen Czerski, broadcaster, physicist and oceanographer at University College London, offers a similar perspective. In her field, contacting the authors of a paper to find out how to replicate their results would be seen as odd. "You might have a chat with them along the way about the difficulties encountered and the issues associated with that research," she said to me, "but you certainly wouldn't ask them what secret method they used to come up with the results." Czerski questions whether Kahneman's proposed rules may breach research ethics. "My gut response is that asking that question is close to scientific misconduct, if you were asking solely for the purpose of increasing your chances of replication, rather than to learn more about the experimental issues associated with that test. The attitude in the fields I've worked in is definitely that you should be able to conduct your own test of a hypothesis, and that you shouldn't need any guidance from the original authors."[17]

One thing seems clear—the culture of replication in the physical sciences is a world apart from the one that prevails in psychology. Katie Mack, astrophysicist at the University of Melbourne, told me that in her field there are many situations where reproducing a result is considered essential for moving the area forward. "A major result produced by only one group, or with only one instrument, is rarely taken as definitive." Mack points out that even findings that have been replicated many times over are valued, such as the Hubble constant, which describes the rate of expansion of the

universe. "Many groups have measured it with many different methods (or in some cases the same method), and each new result is considered noteworthy and definitely publishable."[18]

New rules and constraints on replications—especially when issued by heavyweights such as Kahneman—are likely to discourage academic journals from publishing them, and in this respect journals hardly need further discouragement. With rare exceptions, none of the most prominent journals in psychology or neuroscience regularly publish direct replications, and even the megajournal *PLOS ONE*—one of the few outlets to explicitly disavow the traditional focus on novelty—warns authors that submissions that "replicate or are derivative of existing work will likely be rejected if authors do not provide adequate justification."[19] Populating the literature with false or unconfirmed discoveries is thus deemed acceptable while verifying the replicability of those discoveries demands special explanation.

In contrast to physicists, many senior psychologists seem content with a lack of direct replication. In 2014, Jason Mitchell from Harvard University argued that "hand-wringing over failed replications in social psychology is largely pointless, because unsuccessful experiments have no meaningful scientific value."[20] Mitchell's central thesis was that a failed replication is most likely the result of human error on the part of the researcher attempting the replication rather than the unreliability of the phenomenon originally reported. But in a lapse of logic, Mitchell ignored the possibility that the very same human error can give rise to false discoveries.[21]

Psychologists Wolfgang Stroebe and Fritz Strack have also published a provocative defense of the status quo.[22] Like Kahneman, they argue that in at least some areas of psychology, direct replications are not possible because the original results depend on a host of hidden methodological variables that are just as invisible to the original researchers as to those attempting the replication. Therefore, they argue, when a direct replication fails it is most likely due to the dependence of the effect on hidden moderators rather than the unreliability of the original finding. Instead of defending direct replication, Stroebe and Strack take the view that conceptual replications are more informative because they test the "underlying theoretical construct." They conclude that direct replications are uninteresting regardless of the outcome because they bring us no closer to knowing whether the original study was "a good test of the theory."

Stroebe and Strack's article triggered a trenchant rebuke from psychologist Dan Simons. Simons argues that this characterization of psychology renders it blatantly unscientific because if the failure of a direct replication can always be explained by hidden variables (or moderators) then the original result, by definition, can never be falsified. In such a world there can be no false discoveries. The task of ensuring reproducibility thus becomes impossible because, as Simons puts it:

> the number of possible moderators is infinite: perhaps the effect depends on the phases of the moon, perhaps it only works at a particular longitude and latitude, perhaps it requires subjects who ate a lot of corn as children. . . . We cannot accumulate evidence for the reliability of any effect. Instead, all findings, both positive and negative, can be attributed to moderators unless proven otherwise. And we can never prove otherwise.[23]

In addition, Strack and Stroebe's attack on direct replication can be leveled equally at their preferred alternative of conceptual replication. Whether it succeeds or fails, a conceptual replication could also arise because of the action of hidden moderators or random causes; indeed the fact that a conceptual replication adopts a different methodology can only increase its vulnerability to such confounds. Hidden moderators or (as yet) unexplained causal factors are no doubt a reality in much psychological research, as in other sciences. The rational solution to such complexity is a program of comprehensive direct replication followed by cautious, incremental advances. But discounting direct replication because of the hypothetical actions of unobserved (and unobservable) moderators is tantamount to an argument for magic.

Reason 2: Lack of Power

In chapter 2 we saw how hidden analytic flexibility corrupts science by elevating the rate of false positives. As shown by Joe Simmons and colleagues, exploiting researcher degrees of freedom can have a profound effect on the α level—the probability of erroneously rejecting a true null hypothesis—increasing it from a nominal .05 to a startling .60 or higher. These questionable practices are certainly a major source of unreliable findings in psychol-

ogy. However, an additional and frequently overlooked source is the prospect of experiments missing *true* discoveries. Under the logic of NHST, this is referred to as β: the probability of failing to reject a false null hypothesis. To draw a courtroom analogy, α can be seen as the probability of convicting an innocent defendant, while β is the probability of acquitting a guilty one. The probability of correctly rejecting a false null hypothesis—that is, of correctly convicting a guilty perpetrator—is thus calculated as 1−β. This value is referred to as *statistical power* and tells us the probability that a statistical test will detect an effect of a given size, where that effect truly exists.

Since the 1960s, researchers have known that psychology suffers from low statistical power. Psychologist Jacob Cohen was among the first to draw the issue of power and β into prominence. Cohen surveyed all articles that appeared in 1960 across three issues of the *Journal of Abnormal and Social Psychology*. He found that, on average, they had only a 48 percent chance of detecting effects of medium size and only an 18 percent chance of detecting even smaller effects.[24] Even for large effect sizes, Cohen found that the studies in his cohort bore a 17 percent chance of missing them.

Cohen's analysis of statistical power was groundbreaking and led to a chain of investigations into power throughout the 1960s and 1970s. Even so, these studies had little influence on research practices. When psychologists Peter Sedlmeier and Gerd Gigerenzer returned to the question in 1989 they found that the average statistical power in psychology had scarcely changed between 1960 and 1984.[25] In 2001, Scott Bezeau and Roger Graves took yet another look, repeating the analysis on 66 studies published in the field of clinical neuropsychology between 1998 and 1999. Again, they confirmed an overall power to detect medium effect sizes of about 0.50—virtually identical to previous investigations.[26] For four decades, psychological research has remained steadfastly underpowered, despite widespread awareness of the problem. Most recently, in 2013, Kate Button and colleagues from the University of Bristol extended these findings to neuroscience, uncovering an even lower median power of 0.21.[27]

Why do psychologists neglect power? Psychologist Klaus Fiedler and colleagues have argued that the prevalence of underpowered studies follows from a cultural fixation on α at the expense of β. Doing so, they argue, leads to high rates of false negatives—that is, missed true discoveries—that can cause lasting damage to theory generation. While false positives at least

have the *potential* to be disconfirmed by additional research (albeit minimally, owing to low rates of replication), the unrelenting pressure for researchers to produce publishable positive results means that failed hypotheses are likely to be swiftly discarded and forgotten. Fiedler and colleagues warn that when a correct hypothesis is wrongly rejected and abandoned because of a false negative then "even the strictest tests of the remaining hypotheses can only create an illusion of validity."[28]

Not surprisingly, a common finding from the studies of Cohen onward is that very few published studies in psychology determine sample sizes through a priori (prospective) power analysis. Instead, sample sizes tend to be decided by a rule of thumb in which the sample size is roughly matched to previous experiments that succeeded in obtaining statistically significant effects. The problem with this approach is that it ignores the statistical properties of replication: when a study detects a true positive with a p value just below .05 (e.g., $p = .049$) then an exact replication with the same sample size (assuming the same effect size) has only about a 50 percent chance of successfully repeating the discovery.[29] Increasing the power to 0.8 or higher generally requires researchers to at least double the original sample size. Given the prevalence of "just significant" results in psychology (see chapter 2), when researchers determine sample sizes based on "what worked last time" then the overall statistical power of a field will converge on 0.5 or less. So it is no coincidence that Cohen and others have consistently estimated the power of psychological research to depend on the flip of a coin. Valuing our instincts in sample size selection over and above formal a priori power analysis places a glass ceiling on the statistical sensitivity of any scientific endeavor.

Another reason for the neglect of statistical power in psychology is the almost complete absence of direct replication. Because no two studies employ exactly the same design, this allows researchers to shrug off concerns about power on the grounds that nothing exactly the same has ever been done before, so there is nothing comparable in the literature on which to base a power analysis. This argument is, of course, fallacious for a number of reasons. First, even conceptual replications often apply only small tweaks to previous experimental methods; therefore an estimate of the expected effect size (and hence power) can be gleaned, especially when looking over a body of previous similar work. Second, even in rare cases where power cannot be estimated from previous work, researchers can at least motivate

the sample size to detect an effect of a particular size (e.g., small, medium, or large effects, as defined by Cohen).[30]

It may seem obvious that low power reduces the reliability of psychological research by increasing the rate of false negatives (β), but what about the risk of false positives (α)? Because underpowered studies lack sensitivity to detect true effects, any effects they do reveal must be large in order to achieve statistical significance. Therefore, you might think that a positive result ($p<.05$) from a low-powered study would actually be more convincing than the equivalent finding from a high-powered study—after all, isn't the bar for detecting a true discovery in a low-powered study necessarily higher?

While it is the case that the effect size required to achieve statistical significance is greater for low-powered experiments, this doesn't mean that the probability of those significant results being *true* is necessarily higher. Here we must be careful not to confuse the probability of the data under the null hypothesis (p) with the probability that an obtained positive result is a *true* positive. The probability of a true positive cannot be inferred directly from the p value of an experiment; it must instead be estimated through the *positive predictive value* (PPV). If you imagine traversing a desert in search of water, the PPV can be thought of as the chance that a shimmer on the horizon is an oasis rather than a mirage. Mathematically it is calculated as the number of true positives (oases) divided by the total number of all positive observations (shimmers), both true (oases) and false (mirages):

$$\text{Positive predictive value PPV} = \frac{\text{Number of true positives}}{\text{Total number of positive observations}}$$

To see how the PPV relates to power, we need to substitute these quantities for probabilities. In the numerator, the number of true positives can be replaced by the statistical power of the experiment (1−β) multiplied by the prior probability (R) that the effect being studied is truly positive:

$$\text{PPV} = \frac{(1-\beta) \times \text{R}}{\text{Total number of positive observations}}$$

Then, since the total number of obtained positive results is the sum of all true positives and false positives, we can replace the denominator with

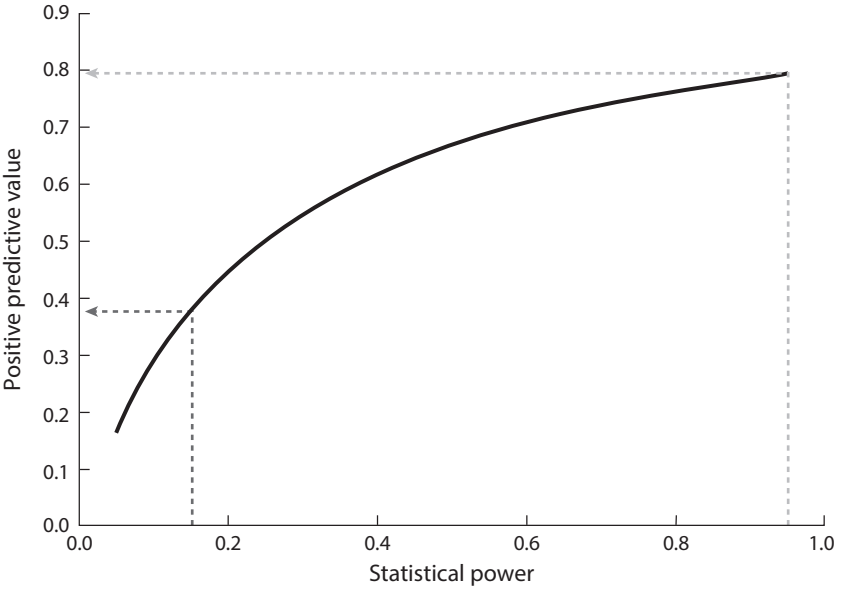


FIGURE 3.2. The statistical power of an experiment tells us its positive predictive value (PPV): the probability that a statistically significant effect is a true positive. This curve shows the relationship between power and PPV for a hypothetical experiment where the prior probability of the null hypothesis being false is 0.2 (i.e., a reasonably unlikely hypothesis) and α =.05. When power is just 0.15 (lower-left dotted line), then the PPV is also just .375. But when power is increased to .95 (upper-right dotted line), the PPV rises to 0.79. This curve plateaus at 0.8; to achieve a higher PPV, the researcher would either need to lower the α or test a more plausible H1.

the probability of obtaining a true positive, $(1-\beta) \times R$, added to the probability of obtaining a false positive (α):

$$\text{Positive predictive value PPV} = \frac{(1-\beta) \times R}{(1-\beta) \times R + \alpha}$$

As we can see in figure 3.2, the PPV is therefore directly related to statistical power: as power increases, so does the chance that a statistically significant effect is a true positive. For example, if we assume a relatively low statistical power of 0.15, combined with the prior probability that the null hypothesis is false of 0.20 (R), then the probability of a positive result being a true positive is only 0.375 (PPV). Increasing the statistical power from 0.15 to 0.95 increases the PPV from 0.375 to 0.79. Therefore, high

power not only helps us limit the rate of false negatives, but it also caps the rate of false positives—just as a powerful experiment increases the chances of finding an oasis, it also reduces the chances of leading us toward a mirage.

As well as failing to recognize the importance of statistical power, psychologists routinely underestimate how complexity in experimental designs reduces power. Psychologist Scott Maxwell from the University of Notre Dame has argued that failure to appreciate the price of complexity could be a major cause of the historically low rates of power in psychology. Even in simple factorial designs, power losses quickly accumulate. Suppose, for instance, you wanted to know the effect of a new cognitive training intervention on weight loss in men and women. To address this question you set up a 2 × 2 factorial design. One factor is the type of training: whether participants receive the new intervention or a control condition (i.e., a placebo). The other factor is the gender of the participants (male or female). Each male and female participant is randomly assigned to one of the treatment groups. Let's further suppose you have key research questions about each of the factors. First, you want to know whether your new intervention works better than the control intervention. This will be the main effect of training type, collapsed across gender. Second, you are interested in whether men or women are generally more successful at losing weight. This will be the main effect of gender, collapsed across training type. And finally, you want to know whether training intervention is more effective in one gender than the other. This will be the interaction of gender × training type. Based on similar studies in this area, you settle on a sample size of 120 participants, with 30 participants in each of the four groups. What would your power be to detect a medium-sized effect for each of the main effects *and* the interaction? Maxwell calculated it to be just 0.47, thus on more than 50 percent of such experiments at least one of your tests would return a false negative.[31]

In 2002, Rachel Smith and colleagues from the University of Michigan reported a series of computer simulations to test the chances of obtaining β errors in experimental designs that have even more factors and explore more complex interactions. What they found was disturbing—at effect sizes and sample sizes commonly observed in psychological research, the rate of false negative conclusions was as high as 84 percent.[32] Their message was clear: researchers ignore power at their peril.

Reason 3: Failure to Disclose Methods

The main purpose of including method sections in published research articles is to provide readers with enough information about the design and analysis to be able to replicate the experiment. A critical reader of any method section should be asking not only whether the reported procedure is sound but also whether it provides sufficient details to be repeatable. Unfortunately, an additional source of unreliability in psychology lies in the systematic failure of studies to disclose sufficient methodological detail to allow exact replication.

Evidence of a systematic lack of disclosure was first uncovered in the survey of 2,000 psychologists by Leslie John and colleagues reported in chapter 2. Based on self-reports, John and colleagues estimated that more than 70 percent of psychologists have failed to report all experimental conditions in a paper, and that virtually all psychologists have failed to fully disclose the experimental measures in a study. In 2013, a team of psychologists led by Etienne LeBel from the University of Western Ontario sought to determine the prevalence of such practices. LeBel and colleagues contacted a random 50 percent of authors who had published in four of the most prestigious psychology journals throughout 2012 and asked whether they had fully disclosed the exclusion of observations, experimental conditions, experimental measures, and the method used to determine final sample size.[33] Of 347 authors contacted, 161 replied—and of these, 11 percent reported that they had failed to report the exclusion of observations or experimental conditions. But even more startling was that 45 percent of respondents admitted failing to disclose all the experimental measures they acquired. The most popular reason for concealing these details was that the excluded measure was "unrelated to the research question."[34] LeBel and colleagues conclude that the time has come for disclosure statements to become a mandatory part of the manuscript submission process at all psychology journals. Disclosure statements could be as simple as the "21 word" solution proposed in the same year by Simmons and colleagues (see chapter 2), which simply requires authors to say: "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study."[35]

As we saw earlier, concealing details about the range of attempted and statistically nonsignificant analyses can dramatically elevate the rate of false positives. At the same time, failing to disclose details of experimental pro-

cedures obstructs other scientists from replicating the work and, in cases where such replications are attempted, sparks additional contention. A case in point returns us to the elderly priming controversy of 2012 between Bargh and Doyen. Recall that Doyen and colleagues failed to replicate Bargh's influential finding that priming people with concepts related to elderly people can lead them to walk more slowly. Doyen claimed that the original result was due to the experimenter in Bargh's original experiments being unblinded to the experimental conditions and inadvertently "priming" the subjects. When Doyen and colleagues repeated the experiment in such a way as to ensure effective blinding, the effect disappeared.

In response, Bargh claimed that the experimenter in his original study *was* blinded to the experimental condition. Unfortunately, it is impossible to tell either way because the method section in Bargh's original study was too vague. Because of this ambiguity, the debate between Bargh and Doyen focused on what the original study did rather than the scientific validity of the elderly priming effect itself—and this dispute spread among the wider psychological community. Such arguments lead nowhere and generate ill feeling: where methodological imprecision is allowed to sully the scientific record, any failed replication can be attacked for not following some unpublished—but apparently critical—detail of the original methodology. And where the original researchers seek to clarify their methodology after the fact, such attempts can be perceived, rightly or wrongly, as face-saving and dishonest.

If lack of methodological disclosure is such a problem, why don't journals simply require more detail? There are no good answers to this question but plenty of bad ones. Space limitations in printed journals require authors to cut details from manuscripts that may be seen by editors as extraneous or unnecessary. Strangely, these policies have persisted even in online (unprinted) journals where space is no concern. Even worse, because direct replication is so rare, much of the detail necessary for replication is regarded as unnecessary by authors, reviewers, and editors. The absence of this detail, in turn, feeds a vicious cycle that makes direct replication of such work difficult or impossible. Methods sections have thus faded from their original purpose of providing the recipe for replication, instead acting as a way for reviewers and editors to check that the authors followed procedural norms and avoided obvious methodological flaws. While this is a necessary role of a method section, it is far from sufficient.

Reason 4: Statistical Fallacies

In chapter 2 we saw how undisclosed analytic flexibility can lead to a range of questionable practices in psychology that undermine reliability, such as *p*-hacking and HARKing. Beneath these errors in the application of statistics, however, lie even deeper and more fundamental misunderstandings of null hypothesis significance testing (NHST).

The most frequently misunderstood aspect of NHST is the *p* value itself.[36] To illustrate just how confusing *p* values can be, consider a scenario where you conduct an experiment testing the effectiveness of a new kind of intervention on the number of people who successfully quit smoking. You give one group of participants the new intervention while the other group is given a control (baseline) intervention. You find that significantly more participants quit following the new intervention compared with the control intervention. The statistical test returns $p = .04$, which at $\alpha = .05$ allows you to reject the null hypothesis.

Now consider which of the following statements is correct:

A. The obtained *p* value of .04 indicates a 4 percent probability that the new intervention was not effective (the null hypothesis).
B. The obtained *p* value of .04 indicates a 4 percent probability that the results are due to random chance rather than the new intervention.
C. The fact that there is a statistically significant difference means that your intervention works and is clinically significant.
D. The obtained *p* value indicates that the observed data would occur only 4 percent of the time if the null hypothesis were true.
E. Previous research found that a similar intervention failed to produce a significant improvement compared the control intervention, returning a significance level of $p = .17$. This indicates that your new intervention is more promising that this previous intervention.
F. After your study is published, another research group publishes a paper reporting that a different intervention also leads to a statistically significant improvement compared to the same control intervention ($p = .001$). Because their *p* value is smaller than your *p* value (.04), they conclude that their treatment is more effective than your treatment.

Which of these statements is correct? In fact, all of them are wrong in different ways. Statement A is the most prevalent misconception of NHST: that a *p* value indicates the probability of the null hypothesis (or any hypothesis) being true. Another common variant of this fallacy is that $p = .04$ reflects a 96 percent chance of that the effect is "real." These statements confuse the *p* value with the posterior probability of the null hypothesis. Recall that NHST estimates the probability of the observed data given the hypothesis, *p*(data|hypothesis), rather than the probability of the hypothesis given the data, *p*(hypothesis|data). To determine the posterior probability of any hypothesis being true we need to use Bayes' theorem, which we will turn to later in this chapter. Statement B is a variant of statement A and equally wrong. For the same reason that a *p* value can't tell us the probability that there is no effect, it can't tell us the probability that the effect is due to chance.[37]

What about statement C? Can we conclude that the new treatment was clinically significant because it was statistically more effective than the control intervention? No, because the *p* value tells us nothing about the *size* of the effect. As statistical power increases, so does the chance of detecting increasingly small effects. Therefore, depending on the statistical power of your experiment, $p = .04$ could indicate an effect so small as to be trivial in terms of clinical significance. Statistical significance and "real world" significance are independent concepts and should not be confused, even though they frequently are.

Statement D moves us closer to the true definition of a *p* value but is still crucially wrong. The *p* value doesn't tell us the probability of observing your data if the null hypothesis is true—it tells us the probability of observing your data *or more extreme data* if the null hypothesis is true.

In statement E, we're asked whether your intervention, which yielded a statistically significant effect ($p = .04$), can be said to be more effective than an alternative intervention that produced a statistically *non*significant effect ($p = .17$). Although this interpretation is tempting—and extremely common—it is incorrect. To conclude that two effects are different requires a statistical test of their difference: in other words, you would need to show that your intervention produced a significantly larger improvement in smoking cessation compared with the alternative intervention. In most cases this requires a test of the statistical interaction, and only if that interaction was statistically significant could you conclude a quantitative dissociation be-

tween effects. In 2011, psychologists Sander Nieuwenhuis, Birte Forstmann, and E. J. Wagenmakers explored the prevalence of this fallacy in 157 articles published in four of the most prestigious journals in neuroscience. Remarkably, they found that 50 percent of papers assumed a meaningful difference between a statistically significant effect and a statistically nonsignificant effect, without testing for the critical interaction that would justify such a conclusion.[38]

Finally, let's consider statement F. This scenario is similar to statement E except that both *p* values are statistically significant. Again, the statement is false without performing a test to compare the magnitude of the improvement between your intervention and the newer intervention. Remember that a *p* value on its own tells us nothing about the size of an observed effect—it merely tells us how surprised we should be to observe that effect, or larger, if the null hypothesis were true. It is entirely possible that a large effect that yields $p = .04$ is more important for theory or applications than a smaller effect from a much larger sample that produces $p = .001$. This tendency for scientists to draw conclusions about effect sizes based on *p* values reflects the same statistical fallacy as statement A: that we tend to interpret *p* values as telling us the likelihood of our hypothesis being true, with smaller *p* values making us feel more confident that we are "right." This thinking is not allowed under NHST because a *p* value can never tell us the probability of truth. Many psychologists (and scientists in related fields) have fallen into the trap of confusing *p* values with a measure that *does* tell us of the relative likelihood of one hypothesis over another: the Bayes factor. We will return to Bayes factors later in this chapter.

Reason 5: Failure to Retract

Scientific publishing is largely a matter of trust, which means that when major errors are identified it is important that untrustworthy papers are retracted promptly from the literature. Articles can be retracted for many reasons, ranging from honest mistakes such as technical errors or failure to reproduce findings, to research misconduct and fraud.[39] If replication is the immune system of science, then retraction can be thought of as the last line of defense—a surgical excision.

Putting aside fraud, in many sciences the failure to replicate a previous result because of technical error or unknown reasons is sufficient grounds for retracting the original paper.[40] In physics, chemistry, and some areas of biology, results are often so clearly positive or negative that failure to replicate indicates a critical mistake with either the original study or the replication attempt. Psychology, however, rarely retracts articles because of replication failures alone. In 2013, researchers Minhua Zhang and Michael Grieneisen found that the social sciences, including psychology, are several times less likely than other sciences to retract articles because of "distrust of data or interpretations."[41] In a previous study they showed that overall retraction rates in psychology are just 27 percent of the average calculated across more than 200 other research areas.[42]

The resistance to retraction in psychology is so hardened that it can lead to farcical interactions between researchers. One recent case highlighted by Dan Simons relates again to the work of Yale psychologist John Bargh.[43] In 2012, Bargh and colleague Idit Shalev published a study claiming that lonelier people prefer warmer baths and showers, thereby compensating for a lack of "social warmth" through physical warmth.[44] In 2014, psychologist Brent Donnellan and colleagues reported a failure to replicate this finding—and not just in a single experiment but across nine experiments and more than 3,000 participants, over 30 times the sample size of the original study.[45] Despite this failure to replicate, as well as the presence of unexplained anomalies in the original data, Bargh and Shalev refused to retract their original paper. In many other sciences, a false discovery of this magnitude would automatically trigger excision of the original work from the scientific record. In psychology, unreliability is business as usual.

What does the rate of retractions in a particular scientific field say about the quality of science in that field? On the one hand you might think (optimistically) that fewer retractions is a good sign, indicating that fewer studies are in *need* of retraction. More realistically, fewer retractions can be thought to indicate lower confidence scientists place in their own methods and a lower bar for publication in the first place. If we take a science such as experimental physics, where studies tend to have high statistical power, methods are well defined and de facto preregistered, then the failure to reproduce a previous result is considered a major cause for concern. But in a weaker science where lax statistical standards and questionable research practices are the norm, attempts to reproduce prior work will often fail, and it should

therefore come as no surprise that retraction is rare. By lowering the bar for publication, we necessarily raise it for retraction. A strict policy requiring retraction of unreliable results could cause the thread of psychological research to unravel, consigning vast swathes of the literature to the scrap heap.

Solutions to Unreliability

What is the point of science if we can't trust what it tells us? Lack of reliability is unquestionably one of the gravest challenges facing psychological science. However there are reasons to be hopeful. In the remainder of this chapter we will consider some solutions to the five main problems outlined above, including lack of replication, low statistical power, lack of disclosure about methods, statistical misunderstandings, and failure to retract flawed studies.

Replication and power. If the problem is a lack of replication and statistical power, then the obvious solution is to produce more high-powered research and instill a culture where, like other sciences, replication is viewed as a foundation of discovery. Our challenge is how to achieve this reform within an academic culture that places little to no value on direct replication, and where jobs and grant funding are awarded based on the quantity of high-impact publications rather than the reliability of the underlying science. Placing a premium on work that can be independently replicated is one major part of the solution. Psychologist Will Gervais from the University of Kentucky has shown how a policy of replication automatically rewards researchers who conduct high-powered studies.[46] Under the current system, a researcher who publishes many underpowered studies ("Dr. Mayfly"; N = 40 participants per experiment) will have a substantial career advantage over a researcher who publishes a smaller number of high-powered studies ("Dr. Elephant"; N = 300 participants per experiment). Using simple statistical modeling, Gervais found that after a six-year period, Dr. Mayfly will have published nearly twice the number of papers as Dr. Elephant. However this advantage is reversed when publication requires every experiment to be directly replicated just once. Both researchers now publish fewer papers, but because Dr. Mayfly wastes so much time and resources chasing down false leads, Dr. Elephant publishes more than

double the number of papers as Dr. Mayfly. By aligning the incentive to publish more papers with the incentive to publish replicable science, the careful scientist is rewarded over the cowboy.

So much for theory, but how can we incentivize replications in practice? E. J. Wagenmakers and Birte Forstmann have proposed an innovative solution in which journal editors issue public calls for replication attempts of studies that hold particular interest or weight in the literature. Teams of scientists would then compete for the bid, with the winning team guaranteed a publication in the journal that issues the call.[47] The journal *Perspectives on Psychological Science* recently launched a similar initiative called Registered Replication Reports in which groups of scientists work collaboratively to directly replicate findings of particular importance, with publication of preregistered protocols guaranteed regardless of the outcome. Sanjay Srivastava has gone even further and called for a "Pottery Barn rule" (you break it, you buy it) in which the journal that publishes any original finding is *required* to publish direct replications of the study, regardless of the outcome.[48] Srivastava's idea is challenging to implement but would incentivize journals to publish only the work they see as credible. It would also capitalize on evidence that replication studies have the most impact when they are published in the same journal as the original paper.[49]

Journal policies must also play a key role in encouraging, rewarding, and normalizing replication. The journal *BMC Psychology*, launched in 2013, is one of a handful journals to explicitly take up the gauntlet, noting in its guide to authors that it will "publish work deemed by peer reviewers to be a coherent and sound addition to scientific knowledge and to put less emphasis on interest levels, provided that the research constitutes a useful contribution to the field." In a provocative article for *BMC Psychology*, editor Keith Laws has noted that this policy explicitly welcomes negative findings and replication studies.[50]

Bayesian hypothesis testing. One of the strangest aspects of conventional statistical tests is that they don't actually tell us what we need to know. When we run an experiment we want to know the probability that our hypothesis is correct, or the probability that the obtained results were due to a genuine effect rather than the play of chance. But the doctrine of NHST doesn't allow such interpretations—the *p* value only tells us how surprised we should be to obtain results at least as extreme if the null hy-

pothesis were true. This counterintuitive reasoning of *p*(data|hypothesis) is why many scientists still harbor basic misunderstandings about the definition of a *p* value.

A more intuitive approach to statistical testing is to consider the opposite logic to NHST and estimate *p*(hypothesis|data)—that is, the probability of the hypothesis in question being true given the data in hand. But to achieve this we need to look beyond NHST—a relatively new invention—and return to an eighteenth-century mathematical law called Bayes' theorem.

Bayes' theorem is a simple rule that allows us to calculate the conditional probability of a particular belief being true (hypothesis) given a particular set of evidence (data) and our prior belief. Specifically, it estimates the posterior probability of a proposition by integrating the observed data with the strength of our prior expectations. Bayes' theorem is expressed formally as:

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

where *P(H|D)* is the posterior probability of the hypothesis given the data, *P(H)* is the prior probability of the hypothesis, *P(D|H)* is the probability of the data given the hypothesis, and *P(D)* is the probability of the data itself.

One of the simplest applications of Bayes' theorem is in medical diagnosis. Suppose you are a doctor who suspects your patient may be suffering from Alzheimer's disease. To examine this possibility you give your patient a 15-minute behavioral test for mild cognitive impairment. The test reveals a positive result, potentially indicative of the disease. So what is the probability that your patient actually has Alzheimer's disease? We can represent this as *P(H|D)*: the posterior probability of the hypothesis that the patient has Alzheimer's disease given that test was positive. Using Bayes' theorem, we can estimate this probability if we have three other pieces of information: the sensitivity of the test, which is the probability that the test would yield a positive result in a patient who actually has Alzheimer's disease, *P(D|H)*, the prior probability of the patient having the disease in the first place, *P(H)*, and the overall probability of a positive test result regardless of whether or not the patient has the disease, *P(D)*. Now, let's further suppose that the test has a sensitivity of 80 percent, which is to say that that it has an 80 percent chance of detecting Alzheimer's disease where the disease is

present; thus $P(D|H) = 0.80$. Let's also assume that the chance of the patient having Alzheimer's disease within the population is 1 percent ($P(H) = 0.01$). To complete the picture, all we need to know now is the chance of the test returning a positive result regardless of whether or not the patient has the disease. To calculate this we can expand $P(D)$ according to the "law of total probability," which proves that the overall probability of a particular occurrence (in this case testing positive for Alzheimer's disease) is the sum of its constituent conditional probabilities:

$$P(D) = P(D|H) \times P(H) + P(D|{\sim}H) \times P({\sim}H)$$

In this equation, $P(D|H)$ and $P(H)$ are defined as above, $P(D|{\sim}H)$ is the probability of a positive result if the patient *doesn't* have Alzheimer's disease (i.e., the false positive rate of the test), and $P({\sim}H)$ is the overall probability that the patient doesn't have Alzheimer's disease (calculated as $1-P(H) = 0.99$). Let's assume that that test in this case has a false positive rate of 5 percent, so that $P(D|{\sim}H) = 0.05$. The overall probability of a positive test result, $P(D)$, can then be calculated as:

$$PD = 0.80 \times 0.01 + 0.05 \times 0.99 = 0.0575$$

We can now substitute this value of $P(D)$ into Bayes' theorem to estimate the probability that the patient truly has Alzheimer's disease:

$$P(H|D) = \frac{0.80 \times 0.01}{0.0575} = .139$$

Are you surprised by this outcome? Despite the fact that the test has 80 percent sensitivity and just a 5 percent false positive rate, Bayes' theorem tells us that there is still only a 13.9 percent chance that a patient with a positive test result has Alzheimer's disease.[51] This outcome violates our common intuitions; indeed, when posed with similar scenarios, even medical practitioners routinely overestimate the accuracy of such diagnostic tests.[52]

Bayes' theorem is a valuable tool in medical screening because it counters our erroneous intuitions about probabilities, but calculating $P(H|D)$ on its own isn't particularly useful in psychological science. In most experimental

settings we need to know more than the absolute probability of a given hypothesis given the data; instead, given the data that we have, we need to know how likely the experimental hypothesis (H1) is relative to the probability of a comparator (baseline) hypothesis. This baseline can be anything we choose, but in general it takes the form of the null hypothesis (H0) that the effect of interest is null or nonexistent in the population being studied. By calculating the ratio between $P(D|H)$ for each of our hypotheses, H1 and H0, Bayes' theorem allows us to decide which hypothesis is better supported by the evidence. The resulting value is known as a Bayes factor, B, and tells us the weight of evidence in favor of H1 over H0.

To calculate B we need to know the relative posterior probabilities of each hypothesis, $P(H|D)$, and their relative prior probabilities $P(H)$. To calculate the relative posterior probability of H1 vs. H0, we simply divide one Bayes' theorem by the other:

$$\frac{P(H1|D)}{P(H0|D)} = \frac{\left(\frac{P(D|H1)\,P(H1)}{P(D)}\right)}{\left(\frac{P(D|H0)\,P(H0)}{P(D)}\right)}$$

This rather cumbersome equation can be restructured as:

$$\frac{P(H1|D)}{P(H0|D)} = \frac{P(D|H1)}{P(D|H0)} \times \frac{P(H1)}{P(H0)} \times \frac{P(D)}{P(D)}$$

Note that the values of $P(D)$ here cancel out to 1, leaving us with the simpler form:

$$\frac{P(H1|D)}{P(H0|D)} = \frac{P(D|H1)}{P(D|H0)} \times \frac{P(H1)}{P(H0)}$$

This equation, in turn, can be restructured to calculate the Bayes factor, B:

$$B = \frac{P(D|H1)}{P(D|H0)} = \frac{P(H1|D)}{P(H0|D)} \div \frac{P(H1)}{P(H0)}$$

In other words, the Bayes factor (B) can be thought of as the ratio of the posterior probabilities, $P(H|D)$, divided by the ratio of the prior probabilities $P(H)$. This provides a simple estimate of how likely one hypothesis is given

the evidence relative to the other, and therefore how much a rational observer should update his or her prior beliefs based on that data. A value of $B = 1$ is perfectly ambiguous and should lead to no change in prior beliefs, while $B>1$ indicates evidence in favor of H1, and $B<1$ indicates evidence in favor of H0. Following the pioneering twentieth-century work of statistician Harold Jeffreys, a value of B between 1 and 3 is generally considered to be inconclusive in favoring H1 over H0, while $B>3$ is judged to be "substantial or moderate evidence," $B>10$ is "strong evidence," $B>30$ is "very strong evidence," and $B>100$ is "decisive evidence." These ratios can be reversed to provide the same standards of evidence in favor of H0 over H1, with $B<0.33$, $B<0.1$, $B<0.033$, and $B<0.01$ providing substantial, strong, very strong, and decisive evidence, respectively.

The rationale of Bayesian hypothesis testing is intuitive, but how exactly does it stand to improve the reliability of psychological science? Recall from chapter 2 that a key source of *p*-hacking in NHST lies in the violation of stopping rules—if we simply add participants to a study until $p<.05$ then we increase the odds that the *p* value will dip below .05 by chance and thus lead to a Type I error. Bayesian analysis, however, is protected against this possibility: according to the likelihood principle the more data we consider, the more accurate our estimation of reality becomes. Thus, Bayesian researchers can continuously add data to their experiments until the evidence clearly supports H1 or H0. This approach not only eliminates the violation of stopping rules as a source of misleading information, it is also allows the most efficient allocation of limited scientific resources.

A second major advantage of Bayesian hypothesis testing is that, in contrast to NHST, it allows direct estimation of the probability of H0 relative to H1. Recall that a nonsignificant *p* value allows the researcher only to fail to reject H0, and never to accept it or estimate its chances of being true. The biased nature of this inference is one major cause of publication bias because any negative finding ($p>.05$) using NHST is by definition inconclusive, which in turn allows authors and journals to feel justified in reporting only positive results. Bayesian inference, however, affords no special status to H0 and permits inferences about the relative probability of any prespecified hypothesis given the evidence: where NHST is limited to concluding an *absence of evidence*, a Bayes factor provides *evidence of absence*. Third, Bayesian tests require authors to transparently specify and justify their prior hypothesis, allowing the scientific community to see just how well grounded

it is, and challenge it as necessary. Finally, Bayes factors can be combined and updated in light of new evidence.[53] Bayesian testing thus encourages scientists to engage in programmatic research that is geared toward replications and accumulated truths rather than the "snapshot science" of NHST where a single study showing $p<.05$ can be impossible to refute.

It must be emphasized that Bayesian hypothesis testing isn't a cure-all for reliability. A Bayesian analysis can still be "*B*-hacked" by manipulating other researcher degrees of freedom such as the selective rejection of outliers or selection of dependent variables. No system of statistical inference is immune to cherry-picking, bias, or fraud. But combined with other initiatives, including study preregistration, Bayesian analysis is a major part of the solution.[54]

Adversarial collaborations. Too often in psychology, as in other sciences, ego overcomes reason, reducing the reliability and credibility of the results we produce. In the competitive world of reputation management, it is often more important to be seen winning an argument than to discover the truth. One proposed solution to this problem is the widespread implementation of so-called adversarial collaborations. An adversarial collaboration is one in which two or more researchers (or teams of researchers) with opposing beliefs jointly formulate a research design that will determine a "winner." A typical formulation involves a team of researchers (proponents) who believe a certain hypothesis to be true, and a counter team (skeptics) who believe that no such effect exists. Having reached a consensus on the optimal design, both sides then implement the experiment and the results are combined at the end to adjudicate between the competing hypotheses. The advantage of this approach is that by agreeing to the rules of engagement in advance (and thus preregistering their hypotheses and analysis plans), both sides are forced to accept the outcome: there is no scope to claim, after the fact, that an undesirable result was the outcome of a suboptimal or flawed methodology used by the opponent.[55]

Improving reporting standards. One obvious way to improve the reliability of psychological science is to ensure that published methods and analyses are more easily reproducible by independent researchers. Recently a number of journals have moved toward more stringent reporting standards. In 2013, the journal *Nature Neuroscience* introduced a new

methods checklist to increase research transparency, while at the same time eliminating words limits on method sections.[56] In 2014, *Psychological Science* introduced a requirement for authors to complete a simple disclosure statement upon the submission of manuscripts:

1) Confirm that (a) the total number of excluded observations and (b) the reasons for making these exclusions have been reported in the Method section(s). [ ] *If no observations were excluded, check here* [ ].
2) Confirm that all independent variables or manipulations, whether successful or failed, have been reported in the Method section(s). [ ] *If there were no independent variables or manipulations, as in the case of correlational research, check here* [ ].
3) Confirm that all dependent variables or measures that were analyzed for this article's target research question have been reported in the Methods section(s). [ ]
4) Confirm that (a) how sample size was determined and (b) your data-collection stopping rule have been reported in the Method section(s) [ ] and provide the page number(s) on which this information appears in your manuscript.[57]

Meanwhile, the Center for Open Science, led by psychologist Brian Nosek from the University of Virginia, has championed an initiative to offer badges for adherence to transparent research practices. These include sharing of primary research data, experimental materials, and study preregistration.[58] Broad uptake and adherence to these new standards remains to be seen, but they are positive and vital moves that emphasize the pressing need to improve reliability and credibility of psychology.

Nudging career incentives. How can we motivate researchers to care more about reproducibility? One approach would be to develop a quantitative metric of reproducibility that counts in career advancement—a system where your score is improved by how often your work is independently replicated. In 2013, Adam Marcus and Ivan Oransky proposed such an index for journals.[59] We will return to this possibility in chapter 8 and consider a concrete mechanism for how it could work in psychological science.

CHAPTER 4

The Sin of Data Hoarding

Code and data or it didn't happen.
—Anon.

If passing aliens were to glance at terrestrial science they might conclude that data sharing is a standard global practice. Scanning the policy of the National Institutes of Health, one of the major US funding agencies, they would see that it "requires resource producers to release primary data along with an initial interpretation . . . to the appropriate public databases as soon as the data is verified. All data will be deposited to public databases . . . and these pre-publication data will be available for all to use."[1] Turning to the journal *Science*, one of the most prestigious peer-reviewed outlets on the planet, they would see that data sets "must be deposited in an approved database, and an accession number or a specific access address must be included in the published paper."[2] They would nod approvingly and move on.

But our visitors should have dug deeper. Had they studied the policies of the NIH and *Science* magazine more closely, they would have discovered that such rules apply only to specific kinds of scientific information, such as DNA and protein sequences, molecular structures, and climate data. Psychology isn't subject to such requirements; on the contrary it is standard practice for psychology researchers to withhold data unless motivated to share out of self-interest or (rarely) when required to share by a higher authority. Unlike fields such as genomics or climatology, many psychologists effectively claim personal ownership over data—data that in the majority of cases is provided by volunteers through the use of public funding and resources.

Many psychologists consider sharing data only where doing so brings professional gains, such as working partnerships that lead to joint authorship of papers. As MIT neuroscientist Earl Miller has said: "I'm for sharing raw data. But as collaborations. Data is not a public water fountain."[3] Many psychologists and neuroscientists share Miller's views. They fear that unfettered access to data would be to surrender intellectual property to rivals, allow undeserving competitors to benefit from our own hard work and invite unwelcome attention from critics. The widespread prevalence of questionable research practices further reduces the incentive to share. After all, in the world of major league science, who would want their mistakes or dubious decision making exposed to the scrutiny of a rival researcher or (worse) a professional statistician? By treating data as personal property and

adopting a defensive agenda, psychology is guilty of our fourth major transgression: the sin of data hoarding.

The Untold Benefits of Data Sharing

Why do some areas of science enjoy a culture of transparency while others, like psychology, are so closed? Are psychologists more selfish than other scientists? Are they more sensitive to competition or criticism? Could it simply be that they are less aware of the gains that data sharing can bring both to wider science and individual scientists?

It is no exaggeration to say that data sharing carries enormous benefits for science. The most immediate plus is that it allows independent scientists to repeat the analyses reported by the authors, verifying that they were performed and reported appropriately. An unbiased perspective can be tremendously helpful in correcting honest mistakes and ensuring the robustness of the scientific record. Sharing study materials (e.g., computer code and stimuli) as well as data can help other scientists repeat experiments, thus aiding in the vital task of direct replication.

Data sharing also makes it easier to detect questionable research practices—were I to conduct 100 statistical tests on a data set and report only the one analysis that was statistically significant (an egregious form of *p*-hacking) then this would become obvious to anyone who tries the 99 analyses on the same data that "didn't work." Beyond the gray area of questionable practices, data sharing also enables the detection of fraud. In 2013, psychologist Uri Simonsohn published a simple and elegant statistical method for detecting fraudulent manipulation of data.[4] As we will see, Simonsohn used this tool on the raw data associated with several published papers and in so doing unmasked multiple cases of data fabrication. His efforts in turn led to university investigations and resignations by established academics.

In the longer term, sharing data and materials at the point of publication prevents the permanent loss of information—over time, computers crash; and scientists change institutions, change contact details, and eventually die, and unshared data inevitably dies with them.[5] This anecdote by psychologists Jelte Wicherts and Marjan Bakker from the University of Amsterdam paints an all-too-familiar picture in psychology: "One of us once

requested data from a close colleague, who responded: 'Sure I will send you those data, but it's like seven computers ago, and so please allow me some time to hunt them down.'"[6]

Finally, publishing data allows independent scientists, now and in the future, to perform analyses that the original authors never imagined or had any interest in pursuing. Complex data sets can hold hidden gems, and aggregated data across many studies can allow researchers to perform vital meta-analysis.[7]

To all these benefits a critic might respond, "Sure, anyone can see how data sharing benefits science, but what about me as an individual researcher? What do *I* stand to gain?" Looking beyond the fact that all publicly funded scientists have an ethical duty to be as open and transparent as possible, one of the most selfish advantages is in citation rates. An analysis of over 10,000 gene expression microarray studies found that those that shared data attracted up to 30 percent more citations than those that did not, after controlling for a range of extraneous factors.[8] Within the field of cancer research this relationship is even stronger, with open data associated with a 69 percent rise in citations.[9] The potential career benefits of data sharing are therefore substantial.

Failure to Share

With so many good reasons to embrace transparency, how many psychologists actually lift the lid on their own data? Public data deposition in psychology is unquestionably rare—very few researchers make their data available to anyone beyond their immediate research team. Nevertheless, all major psychology journals require authors to retain data for several years and share it with readers on request. This raises the question of how many psychologists meet these requirements and share data when asked.

In 2006, Jelte Wicherts and colleagues decided to put this policy to the test by requesting data from 249 studies that appeared in four major journals published by the American Psychological Association (APA). Their experience was unsettling:

> Unfortunately, 6 months later, after writing more than 400 e-mails—and sending some corresponding authors detailed descriptions of our

> study aims, approvals of our ethical committee, signed assurances not to share data with others, and even our full resumes—we ended up with a meager 38 positive reactions and the actual data sets from 64 studies (25.7% of the total number of 249 data sets). This means that 73% of the authors did not share their data.[10]

In the 73 percent of cases where data wasn't obtained, the original researchers either failed to respond to e-mails (18 percent), refused to share the data (35 percent), or promised to send it but never did (20 percent). This was despite the fact that all the studies were published within the prior 18 months, and despite clear APA guidelines requiring the authors to retain data and provide it to interested readers.[11] In a separate interview, Wicherts told me that among the specific reasons for refusing were statements such as, "I am up for tenure and I have no time to do this"; "This is an idiotic amount of work. I am sorry but I can't help you"; "This is an ongoing project"; and "I am afraid your request is not possible."[12]

As a solution, Wicherts and colleagues called for public data deposition to become a mandatory condition for publication. A decade later, the APA is still yet to act on their recommendation, and only a few major psychology journals have issued a blanket rule requiring public data deposition.

One reason why psychologists might refuse to share data is for fear of their results being disputed owing to questionable or sloppy research practices. If this is the case then statistical errors, particularly those favoring statistical significance, should be more common in studies where researchers fail to share. In 2011, Wicherts and colleagues tackled this question by requesting the raw data for 49 studies published by two APA journals: the *Journal of Personality and Social Psychology* and the *Journal of Experimental Psychology: Learning, Memory and Cognition*.[13] Consistent with their earlier investigation, they found that fewer than half of the authors (43 percent) were willing and able to share at least some data. They also found that misreporting of *p* values was more than twice as frequent in studies where the authors failed to provide the data. Most of these errors were clear but relatively minor—more worrying were the 10 cases where the correct calculation of *p* values, based on details in the published article, transformed a result from being statistically significant to statistically *non*significant. For all 10 of these errors, across 7 studies, the authors were unable or unwilling to produce the data supporting their conclusions.

In the face of such evidence you might think that the APA would, at a minimum, reprimand authors who fail to share data and even retract the offending papers. Not so. Despite violating the APA's code of conduct, none of the cases documented by Wicherts and colleagues has met with official sanction or criticism of any kind. Wicherts reflects, "The APA did not pursue any specific authors for not sharing data with us. I am also not aware of any case in which the failure to share data has led to a formal investigation or sanction in psychology. In fact, if a given author refuses to share data, I fear that there is little you can do. In the last 8 years, the APA has done nothing to improve data availability in any of its 50+ journals."[14]

Thus, while the APA code pays lip service to data sharing, it apparently does nothing to enforce it. Few psychologists, and least of all the APA, seem to care whether psychologists share their data or not.

Secret Sharing

Even when psychologists are compelled to share data, the ability of other researchers to work with it can be hampered by usage restrictions. One such recent case returns us to the 2012 study by John Bargh and Idit Shalev discussed in chapter 3.[15] Recall that in their high-profile paper, they claimed that lonelier people prefer warmer baths and showers. Their explanation for this curious correlation was that lonely people receive less social warmth and so try to make up the difference by increasing their exposure to physical warmth.[16] In 2014, Brent Donnellan and colleagues attempted to replicate Bargh and Shalev's results in a much larger sample and found no such relationship. To try to understand the basis of Bargh and Shalev's claims they requested the raw data of the original study, which was published in the APA journal *Emotion*, and so was subject to the following clause in the APA code of conduct:

> After research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis.[17]

According to the APA policy, Bargh should have shared the data with his fellow professionals. Donnellan explains how Bargh shared the data with

Donnellan's colleague, Joe Cesario, but only with extraordinary conditions placed on how it could be used:

> Bargh let Joe Cesario look at his data but he forbids us from talking about what Joe observed. So a gag order is in place. I think this is bull****. There is no reason why there should be a veil of secrecy around raw data. How can we have an open and transparent science if researchers are not allowed to make observations about the underlying data used to make published claims?[18]

Psychologist Dan Simons, who had previously spoken up against Bargh following the replication debacle on elderly priming (see chapter 1), also expressed concern about these restrictions:

> As a field, we should be concerned whenever a scientist imposes a gag order on the public discussion of their data from published studies, strikes out against those who, for whatever reason, can't replicate results, or deletes public critiques of his arguments. Secrecy without explanation or justification is antithetical to the open discourse central to the scientific process. Openness is essential to seeking out the truth, and secrecy about published results leaves outsiders to wonder about the reason for all the fog.[19]

It seems unlikely that Bargh's restrictions on the use of his data would be legally enforceable. Nevertheless, Donnellan and colleagues respected the gag order out of professional courtesy. That researchers should feel entitled to such secrecy is invidious but unsurprising within a culture where data is regarded as the personal property of the researcher who collected it.

How Failing to Share Hides Misconduct

The most unscrupulous "researchers" of all, fraudsters, have every reason to be wary of sharing data. In 2011 and 2012, Uri Simonsohn undertook a remarkable investigation to expose two major cases of data fabrication using statistical methods alone—cases that led to him being popularly dubbed the "data detective."[20] The first case was triggered by an article on social embodiment in the *Journal of Experimental Social Psychology*, led by re-

searcher Lawrence Sanna. Sanna, from the University of Michigan, claimed that participants placed in higher physical locations, such as at the top of escalators, behaved more generously toward strangers than participants at lower elevations—this was argued to be consistent with a theorized relationship between morality and height (as reflected in common phrases such as "the moral high ground"). Over three studies Sanna claimed that people in higher locations spent more time helping others, were more compassionate, and behaved more charitably.

Simonsohn's suspicions were roused when he noticed an odd pattern in the results. While the reported averages in each study differed substantially between the different elevation conditions, the standard deviations were almost the same. For example, in one of Sanna's experiments, the average amount of hot sauce given by research participants to strangers in a taste test (a common proxy for how compassionate participants are to strangers) was reported to be 39.74 grams in the "high" condition, 85.74 grams in the "low" condition, and 65.73 grams in the "control" condition. In contrast, the standard deviations for each condition were virtually identical at 25.09, 24.58, and 25.65, respectively. So while the averages more than doubled between conditions the standard deviations of those averages varied by just one-fiftieth of the same amount. This led Simonsohn to suspect that the data were too clean to be the result of random sampling.

Of course, just because data look too good to be true doesn't mean they are the product of fraud. Any number of more benign explanations must first be ruled out, such as questionable research practices in which the authors report only a subset of experiments that showed the clearest effects or fail to report otherwise legitimate data exclusions. The authors in such cases might even be entirely blameless. Given the strong bias of journals to publish clean and striking results (see chapter 1), their article could simply be the one out of thousands that happened to "strike it lucky"—the so-called winner's curse.

To test these possibilities, Simonsohn studied other articles by the same authors as well as articles that used the same experimental paradigms by different authors. The pattern he found was worrying. Several articles authored by Sanna showed the same similarity between standard deviations, and this similarity was much greater for these articles than for a sample of similar papers in the same field by *other* authors. Simonsohn then requested

the raw data from Sanna (which, perhaps surprisingly, he provided), and using simulations based on the actual distributions of the data he estimated the probability of such patterns occurring due to chance to be 1 in several billion. On this basis, Simonsohn alerted the relevant university authorities. Sanna's articles were subsequently retracted at his own request, and he resigned.

Using the same methodology, Simonsohn also exposed another case by Dirk Smeesters from Erasmus University in the Netherlands. Not only did Smeesters report standard deviations that were too similar to be obtained from random samples, he also left another telltale sign of fraudulent intervention—a human bias stemming from the *law of small numbers*. When people are asked to generate a short sequence of random numbers, say between 1 and 10, they typically overestimate the spread of the data that would arise by chance. So, for instance, in a sequence of a dozen numbers between 1 and 10, a typical sequence produced by a person might consist of 1, 3, 4, 4, 5, 6, 7, 7, 7, 8, 9, 10. Truly random sequences, however, can often involve seemingly unlikely repetitions, such as 1, 2, 3, 4, 4, 4, 4, 4, 6, 7, 8, 8. When a human fabricates data they will typically obey the law of small numbers and overestimate the spread, which can be measured by studying the frequency of the most commonly occurring number (the mode). For the humanly generated sequence above the mode is 7 (with a frequency of 3), while for the randomly generated sequence the mode is 4 (with a frequency of 5). Using simulations based on raw data provided by Smeesters, Simonsohn found that the frequency of the mode was consistently too low to be due to random sampling—it occurred just 21 times in 100,000 simulated experiments, suggesting that it was generated artificially. Following a university investigation, several of Smeesters's articles were retracted, and he resigned even before the investigation had concluded.

The most striking feature of these cases is that while the aggregate data that is typically reported in journal articles provided the initial suspicion of unreliable results, it was the systematic analysis of multiple raw data sets that provided the smoking gun of data fabrication. In the words of Haldane, "Man is an orderly animal. He finds it very hard to imitate the disorder of nature."[21] Only by scrutinizing raw data can the scale of such imitation be uncovered, and only by introducing a culture where data sharing is the norm can such cases be efficiently detected and purged from the scientific record.

Making Data Sharing the Norm

Compulsory data sharing may be ideal, but the road to shifting norms can be rocky, as the PLOS family of journals discovered. Since March 2014 all PLOS journals now "require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception. . . . Refusal to share data and related metadata and methods in accordance with this policy will be grounds for rejection."[22] Among the acceptable methods for data sharing are: deposition in a public archive, making the data available via a third party, or attached as supplementary files that are stored and made available by the journal. Recognizing the impotence of journal policies that require authors to share data on request, PLOS took the bold step of mandating the release of data as a condition of publication.

The reaction from some corners of the scientific community was swift and scathing. "PLOS is letting the inmates run the asylum and this will kill them," declared prominent science blogger DrugMonkey.[23] Among a long list of objections, DrugMonkey pointed to the increased personnel time required to prepare data for public release, a fear that mandatory sharing could lead to "data orthodoxy" in which rigid and unrealistic standards for data curation are imposed, and the worry that publicly available datasets could fuel nitpicking by science denialists. Other critics have also criticized the infeasibility of uploading very large datasets (such as terabytes of video files), and the fact that data sharing may not in any case prevent or uncover the most sophisticated acts of fraud.[24] Some researchers further worried that repeated statistical testing within public data archives could increase the rate of false discoveries, that sharing could breach the confidentiality of human participants, and that it could encourage acts of "scooping" in which rivals scavenge archived data and publish on it before the authors who produced it have gained the benefits they deserve. It has even been suggested that mandated sharing could disadvantage third world countries by forcing them to release their most precious commodity to the slick machinations of first world labs.[25] Under sustained attack, the PLOS data-sharing policy soon earned an ignominious hashtag on Twitter, #PLOSFail.

The intense criticism of PLOS in turn spurred some robust defenses of open science. Writing on his blog, geneticist Matthew MacManes of the University of Hampshire accused DrugMonkey and other bloggers of adopt-

ing "myopic" and "laughable" positions. "If you are working in academia, in the US, you are very likely to be funded by taxpayer money," MacManes said. "The data you produce is enabled by taxpayers. . . . So why is it such a hard friggin' concept that you should be required to share your data freely, upon publication?"[26] MacManes argued that in most cases it is easy to share data—concerns that doing so is too hard are "corner cases" that are unrepresentative of most science published in PLOS. He left readers with a challenge: "Go to any of the PLOS journals and look at the last month of publications. How many of these contain data so super-special or large that they could not be posted?"

Viewed individually, none of the objections to the PLOS policy seem to present a roadblock to sharing. For example, if we consider fears about false discovery rates due to "data fishing" in online archives, this concern is easily addressed by highlighting the provenance of the data, recognizing—with due caution—that analyses of existing archives should be seen as acts of hypothesis *generation* that catalyze future hypothesis testing, and never as hypothesis testing per se. Concerns about confidentiality can also be met, in most cases, through careful anonymization, while concerns about scooping can be solved by archiving some data sets with a future embargo date.[27] These so-called dark archives can provide original authors with a head start of months or even years over their competitors.

While many aspects of the PLOS backlash are addressable, others highlight how existing mechanisms for open data need to be streamlined. Some of the most frustrating problems are trivially technical. Many of the available (free) data-archiving services, such as Figshare, Zenodo and Dataverse, place a limit on the maximum size of individual files that can be uploaded even though they don't limit the overall size of an uploaded archive.[28] In some areas of psychology very large files are routinely acquired, such as videos or neuroimaging data sets, and arbitrary file size limits are an unnecessary deterrent to sharing. Even with such barriers removed, there is no escaping the fact that archiving data and code in a form that others can comprehend requires a much greater investment of time than researchers are accustomed to spending. Among all the commodities in science, staff time is among the most expensive, so a commitment to data sharing will need to be recognized and supported by both funders and academic institutions.

The PLOS policy itself has teething problems. One key question is what exactly constitutes raw data and how adherence to a sharing policy can

be policed within the sprawling architecture of a megajournal. Unfortunately, concerns about enforcement appear to have merit. In writing this book I drew a random sample of 50 studies employing brain-imaging methods between 1 November 2014 and 1 May 2015, and it was remarkable how few researchers shared any data at all.[29] In spite of the policy coming into effect months earlier, just 38 percent of papers in the sample archived their raw fMRI data for restricted or unrestricted access. Authors who failed to share data consistently cited legal or ethical restrictions, yet there was also a great deal of inconsistency in the application of such exemptions; while some authors claimed to be unable to share data with anyone, others from the same countries were able to share comparable data without restriction.

Legal and ethical restrictions are one thing—and in many cases may be justified—but the most worrying trend in this sample of PLOS articles was how often authors appeared to simply sidestep the policy. In 25 percent of papers (13 out of 50), the authors claimed that all study data was present in the manuscript or the supporting information when it is obvious that neither was the case. In place of data, the paper and supporting information instead contained summaries of aggregate data, such as figures or graphics. Somehow—though it is difficult to conceive how—such studies survived the review process and were considered by academic editors to be in compliance with the PLOS data-sharing policy.

In June 2015, I raised these concerns with PLOS executive editor Damian Pattinson. Following an internal investigation PLOS told me that, as of December 2015, most of the 13 cases had now shared at least some data, and that the journal was working on updated guidelines to help avoid future cases:

> The number of instances of authors making their full underlying data available has risen dramatically since PLOS introduced its data mandate. Staff carry out a number of checks to submissions in relation to our data requirements and we support our Academic Editors if they identify any concerns, since they are best placed to advise on whether the dataset is complete and adequately pertains to the findings reported in the manuscript. However we acknowledge that we do not yet have full compliance on *PLOS ONE*.
>
> If, upon publication of an article, it transpires that the data are not fully available, we follow up with the authors to request deposition. We

> have done this in the cases raised here and so far data for 10 out of the 13 studies have been deposited, and this information has been made available to readers via a comment on the articles. We are pursuing our follow up for the remaining three articles. We will continue to pursue cases that do not adhere to our policies, and are developing more subject-specific guidelines for authors so that it is clear what data are required and how they should be deposited.[30]

Compared with PLOS, specialist psychology journals have been slower to take up the mantle of open data—of the dozens of available journals, very few require data sharing as a condition of publication. Nevertheless there is evidence of slow but steady evolution. Recent years have seen data-sharing policies emerge at the journals *Cognition*, *Experimental Psychology*, and the newly established APA journal *Archives of Scientific Psychology*.[31] Each of these requires data deposition at the point of publication, although *Cognition* and *Experimental Psychology* permit exceptions for legal, ethical, or practical reasons. *Archives of Scientific Psychology* offers a stronger policy, mandating data sharing for all articles while also stipulating that "Next users [of the data] agree to offer authorship to the originators of the data on any subsequent publications." Whether providing data alone is a sufficient contribution to warrant coauthorship is controversial, and just how such a policy of mandatory coauthorship can be enforced remains to be seen.[32] Nevertheless, encouraging collaborative relationships between sharers and users offers a clear incentive for authors to adopt open practices.

Short of mandating sharing, there is also evidence that so-called nudge incentives increase open practices. Since 2014, *Psychological Science* has published open data kitemarks known as "badges" alongside articles that share data freely in a public repository.[33] Since the implementation of open data badges, the percentage of articles sharing data has risen steadily from less than 5 percent in 2013 to 25 percent in 2014 to 38 percent in 2015. In June 2015, the journal *Science* went one step further in publishing the Transparency and Openness Promotion (TOP) guidelines, which like the badges initiative is the brainchild of Brian Nosek and colleagues at the Center for Open Science.[34] The TOP guidelines, which I helped coauthor, call for journals to periodically review their adherence to various levels of openness, including standards for data sharing. Although the guidelines do not mandate adherence to transparent practices, they raise public awareness of the journal's adherence to transparency. At the time of writing, the guidelines

have attracted over 750 journals and 60 major organizations as signatories, though notably not the APA.[35]

Top-down reforms are being driven not only by journals. In the UK, the major research councils are increasing their commitment to ensure openness and long-term preservation of data. The Economic and Social Research Council (ESRC), which funds more psychology research than any other agency, now requires grant holders to deposit data in permanent archives.[36] In contrast to the PLOS policy, the ESRC's directive met with no objections. Psychologists, it seems, have learned to not bite the hand that feeds them.

Grassroots, Carrots, and Sticks

Changes in policy are no doubt effective in changing practices, but they can also be slow to take shape. Gatekeeping organizations, such as journals and funders, are often managed by complex and conservative bureaucracies that instinctively (and sometimes mindlessly) constrict toward the status quo. A more radical alternative to driving reform, proposed by Cardiff University psychologist Richard Morey, is to harness the power of gatekeepers at a grassroots level—the level of peer review.

Science depends crucially on the unpaid service of thousands of *peer reviewers*: researchers who in most cases donate their time freely to critique the work of other researchers in the assessment for publication. Collectively, reviewers wield enormous control over science and the careers of individual scientists—they decide which papers meet a sufficient standard for publication, assessing whether the scientific question tackled by a study is well justified, whether the methods are valid, and whether the conclusions of the authors are based on the evidence. In 2014, Morey and colleagues (including me) launched a project called the Peer Reviewers' Openness (PRO) initiative, which invites reviewers to add one more item to the list of essential manuscript ingredients: that it should share data in a public archive or provide a reason for not sharing.[37] From 1 January 2017, signatories to the PRO initiative pledge not to review papers in which authors both refuse to share data *and* also refuse to say why they cannot share. In this way, signatories can withhold reviews just as any reviewer would refuse a paper that lacked critical components such as a method or results section.

The PRO initiative attempts to shift norms in a unique and (potentially) very powerful way. Rather than lobbying journals and funders to change

their policies, it harnesses the academic freedom reserved by all individuals in the scientific community—the choice of each researcher to decide that the publication of data, where possible, is an integral part of the scientific process. As more reviewers sign up to PRO, the incentive for researchers to embrace open practices will, in turn, increase. Authors who choose not to share (and refuse to justify their closed practices) will struggle to find reviewers to assess their manuscripts, delaying their publication workflow and placing them at a career disadvantage compared with those who do share or are able to justify not sharing. At the same time, within a peer-review system that is already heavily burdened, journal editors will face mounting pressure to implement open data policies in an effort to ensure a steady supply of reviews. PRO also carries a direct reward for reviewers themselves. Because data deposition is rare in psychology, reviewers who sign up will find their review workloads substantially reduced, at least in the short term.

When PRO was announced in November 2014 (under the now superseded name "Agenda for Open Research") it was met with a mixture of enthusiasm and trepidation from the psychological community. On the Facebook group of the International Social Cognition Network (ISCON), the initiative was criticized by some psychologists as divisive and heavy-handed. In a heated debate, one senior psychologist warned that the agenda was "a brilliant move, but coercive and politicizing." He said, "I am sympathetic to the cause now, but am concerned with such mass 'work stoppage' in order to gain concessions from the field. . . . An academic boycott is a dangerous thing and something we have thankfully not seen in our discipline . . . till now. Are we really prepared to go down this route?" Another researcher likened the initiative to the academic boycott of Israel and called instead for "a host of less coercive methods." He added, "We can hold symposia on the blessings of data sharing. We write articles and books about this topic. The sky is the limit. . . . My ideal of open science is a science that persuades, not a science that forces people to do things against their will."

In response, Richard Morey argued that the initiative doesn't force authors to share their data, only to say why they refuse to share. "Since you can meet the requirements of the agenda simply by giving an explicit reason why your data is not open, and then the reviewer is asked to take your word for it, I can't understand why so many people here have a problem with it. All you have to do is explain why your level of openness is reasonable. Why is this unreasonable?"

The answer, of course, is that the real reasons authors decide against sharing are at a minimum embarrassing and at worst professionally imperiling. Were psychologists to answer honestly, the (tongue-in-cheek) truth would be:

- "I'm not sharing my data because I can't be bothered to organize the files in a way that makes sense to anyone else—in fact, they probably won't even make sense to me when I look back at them in six months' time."
- "The last thing I need is a bunch of second-stringers combing through my raw data to look for 'mistakes' in my analysis. And, believe me, given how little time I spent rechecking my analysis (once I'd found what I was looking for) there are bound be some mistakes!"
- "Share data? How would I explain the three dependent measures I strategically omitted from my report because the findings didn't fit with my conclusion? Or the five equally defensible ways of analyzing the data that produced completely different outcomes?"
- "The importance of my data is dwarfed only by my towering professional reputation. Were I to publish my data, hordes of competitors would descend like vultures, publishing on my dime and depriving me of dozens of rightful *Nature* and *Science* papers. I'm a scientist, not a charity."
- "You only want me to release my study materials and code so that you can benefit from all my hard work. To which I say, go and write your own code!" (shakes fist)
- "I can't possibly share my code because firstly it isn't commented, secondly it is ugly and nonsensical, and fourthly was rushed and probably missing something."
- "Sharing my experimental materials would only make it easier for other scientists to fail to replicate my results. As things stand now, when someone tries to replicate me using my vaguely worded method sections, I can at least shrug away their nonreplication as due to incompetence on their part, or some unstated but (obviously) crucial difference between their study and mine. The moment I start sharing materials, failed replications might suggest that my work isn't reproducible!"

- "I'm not sharing because it would be a huge amount of work for no immediate reward. This whole 'transparency movement' is already making it harder for me to spin the stories I need to get my papers published in the *Prestigious Journal of Implausible Results*. The last thing we should be doing is encouraging the crusaders."

Unlocking the Black Box

Some years ago, as a novice journal editor handling one of my first assignments, I saw for myself how closed practices hinder the advancement of science. A manuscript I was editing came back with two quite critical, but overall positive, reviews. In one instance, the reviewer chose to reveal her identity to the authors because the authors had claimed that a result shown in one of the reviewer's earlier studies, and which conflicted with the authors' current findings, might not hold up under a reanalysis using a different method. The reviewer responded by arguing that this wasn't the case, given her intimate knowledge of her own data, and so asked the authors to modify their claim. In response, the authors requested the relevant data from the reviewer so that they could run the analysis for themselves. Remarkably—or perhaps *un*remarkably given what we now know about standard practices in psychology—the reviewer denied the request on the grounds that "I usually do not share the raw data of my experiments with people that I'm not collaborating with." This placed the authors (and me, as editor) in a difficult situation. Here was a reviewer demanding a revision of a manuscript based on privileged information that only the reviewer held and was unwilling to disclose. Given the reviewer's refusal to be transparent, I followed what I felt was the only defensible course of action: I disregarded the reviewer's concern. To this day, nobody knows what an independent reanalysis of that data might have revealed. I have no idea who benefits from such a stalemate, but it certainly isn't science.

Years later, such stories remain common in psychology, but a combination of initiatives is shifting the field steadily in the right direction. As journals and funders strengthen their data-sharing policies (despite the teething problems at PLOS), the TOP guidelines are raising the profile of those policies and generating vital discussions. Meanwhile the PRO initiative is incentivizing open practices at a community level. But even beyond

carrots and sticks, there are also a growing number of researchers leading by example and embracing open practices for no reward. This raises the question of why, in the current system, a psychologist would choose to push against the grain, and what can we learn from such acts of apparent altruism?

To find out I asked psychologist Frederick Verbruggen from the University of Exeter about his experiences and opinions on open data. Along with Jelte Wicherts, Verbruggen is one of a handful of psychologists blazing the trail on data sharing—since 2013, all publications arising from studies in the Verbruggen laboratory are accompanied by deposited data in a freely accessible public archive. Why has Verbruggen taken the step to embrace open practices? He told me it was a gradual evolution precipitated by an incident in which he attempted to reanalyze some existing data from a study he ran several years earlier. "As a consequence of not having a decent data management plan back then, I now found myself really struggling to find the appropriate data sets (and lost a lot of time). This was not the first time that had happened to me, so I finally realised that I had to change my data storage practices."[38] Verbruggen and data officer Myriam Mertens then set about developing a series of data management and sharing guidelines in his laboratory. "Once we started implementing the guidelines, I realised that it was incredibly straightforward to share the data for individual studies. This allows reviewers to check our data or analysis protocols if they want to. Once the paper is published, other researchers are free to check and use our data as well."

Verbruggen admits that the transition from closed to open practices was hard, but nowhere near insurmountable. "At the beginning I had to change how I was storing data, and this did take some effort. But once the systems were in place, it all became very straightforward. If you have data sharing in mind from the moment you start programming your experiments, you can take a few small steps that make data sharing at the end very easy. This includes simple things, such as adding a documentation file that briefly describes the experimental procedure and what you can find in the data files." Now that such practices are established in his research he believes they carry personal benefits in addition to the gains for other scientists. "I hope that people will trust my work more because I am completely open about how we analyse our data now." His lab releases its analysis scripts in the (free) statistical software package R, allowing readers to trace every step

of the analysis. Doing so, he argues, "may also increase citations and visibility of the research." He also points out the praise that sharing attracts from colleagues. "So far, several reviewers have commented positively on my data sharing practices. I don't think it has influenced their overall rating of our papers, but it is nice to see that the extra effort is appreciated. I also provide a link to data at the end of my presentations, and some people have commented positively on that as well."

Verbruggen is enthusiastic about sharing. "I would strongly recommend it to everybody," he says. At the same time, he also believes in the importance of authors receiving credit for sharing data. "Others can freely use them, but I do expect people to cite the source of these data. For example, we upload everything on our institutional repository, which provides a handle, and for every study we add a statement that says 'Unrestricted use permitted but please acknowledge source and include the dataset handle.' For me this is a matter of scientific integrity."

What does Verbruggen make of the storm of criticisms thrown at data-sharing policies, such as those launched by PLOS? Although strongly supportive of data sharing at the point of publication, Verbruggen is emphatic about the importance of opt-out mechanisms where sharing may be unethical or infeasible. "I'm the first to admit that data sharing is rather easy for purely behavioural studies, whereas it may be less straightforward for neuroscience studies with large data sets. But even for those studies, it should be possible to share at least some aspects of the data." He also recognizes the concerns of researchers who fear being scooped or prefer to share data in exchange for authorship. "Some studies take a very long time to run and lead to large, multifaceted data sets. And indeed, in some sub-disciplines, it is common and perfectly acceptable for researchers to extract multiple papers from such data sets. So I can understand that the researchers want to have first dibs, or at least a co-authorship, in such situations." Is the solution in such cases not to share? No—instead Verbruggen advocates the use of embargoed data sets (or so-called dark archives). "You can impose an embargo, so that only after a specified period the data become publicly available. Another option is that the owner of the dark archive data has to give permission for each request for access. The advantage is that data can still be preserved for the long term, while owners retain control over data re-use."

One point is clear. Even among its strongest advocates, data sharing in psychology is far from a black-and-white issue. Dorothy Bishop has warned

that the public availability of large multivariate datasets could encourage a form of *p*-hacking in which researchers trawl through archived data to search for patterns in support of particular conclusions—and in some cases, such analyses could even be weaponized in order to discredit the original research (or researchers) in pursuit of a political objective. Bishop also highlights the dangers of publication bias: "Suppose we have a researcher interested in the possible impact of diet on children's cognitive development. There are several large publicly available datasets that could be used to address this question. If a positive association is found, then the result is reported, but if it is not, the researcher is likely to move on to look at something else—unless they embarked on the analysis with the intent of disproving the link."[39] As a solution, Bishop suggests that secondary analyses with clear hypotheses should be transparently preregistered, with unregistered analyses considered exploratory.

Verbruggen stresses that he would never seek to impose a policy of mandatory or universal data sharing blind to the consequences. Instead he appeals to the better angels of our nature, noting simply that he "would just strongly encourage people to share their data whenever reasonably possible, which should be possible for the vast majority of cognitive psychology and basic science studies." In psychology it appears that the devil lies in negotiating a consensus in what different researchers consider to be *reasonable* and *possible* sharing of their chief currency. When asked directly, most researchers agree that transparency and data sharing are good (even best) professional practice. However, with few incentives to share—and several incentives not to—reform may be a long road. For data sharing to succeed in psychology, the community needs to see personal gains. As one Twitter user, Greg Hale, put it: "I'll trade access to my 2 TB [terabytes] for the world's exabytes any day!"[40]

Preventing Bad Habits

Verbruggen successfully reformed his lab practices, but what about preventing suboptimal practices *before* they can take hold? Psychologists Candice and Richard Morey have proposed an innovative "data partners scheme" in which early-career scientists in different labs pair up in order to swap data, develop a five-year data curation plan, and check that each other's data can

be understood before submitting to an archive. The overall aim of the scheme is to steer young scientists toward open practices from the earliest point, instilling awareness of the importance of openness and the necessary skill set to achieve it:

> For many scientists, open science is a difficult choice because it is encumbered by a number of unnecessary pragmatic concerns flowing from habits formed over many years. Openness is not truly a free choice, driven by the merits of open science. This need not be the case for the next generation of researchers. Senior researchers have an important role to play in helping their advisees form good habits and develop a theory of scientific mind. The data partner scheme, the five-year plan, and the submission check can help establish good lab practices, with the benefit that students will be prepared for a more open science.[41]

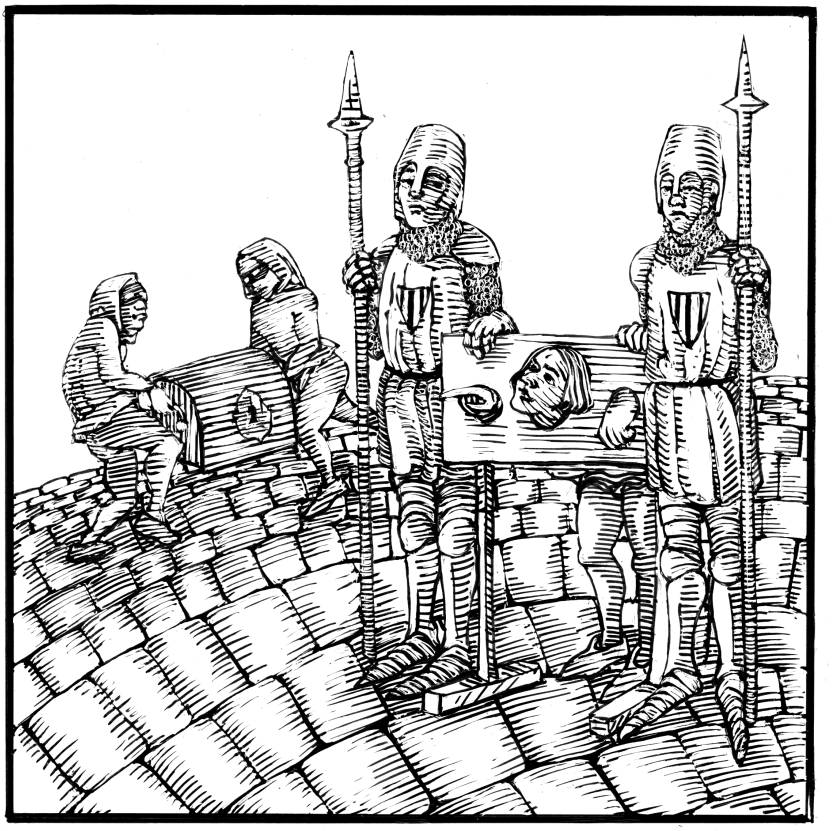
Other sciences show us what we can achieve when we reach the end of that road. In crystallography, the definitions of what constitute reasonable and possible data sharing are now unanimous, and the field has embraced transparency. Stephen Curry, professor of structural biology at Imperial College London, believes that data deposition has strengthened the quality of science in his field. "Publication now requires deposition of the coordinates of the published structure in the Protein Data Bank and I think I'm right in saying that the X-ray diffraction data on which the structure is based should also be deposited. This helps to encourage healthier attitudes to data analysis and should enable improved future analysis as better software tools become available. The crystallographic community is generally quite proud of its openness in this regard."[42] With much effort and a little bit of luck, psychology may one day be able to say the same.

CHAPTER 5

The Sin of Corruptibility

I was doing fine, but then I became impatient, overambitious, reckless.

—Diederik Stapel, 2012



It's late. Everyone has gone home, and you're tired. After a grueling week you have finally collected all the data for your latest study. It's a crucial experiment, one that your manuscript reviewers demanded in order to corroborate the four original experiments in your paper. If the results look good, your work will most likely be accepted for publication in *Nature*. This would be a "career maker," a phrase your supervisor says with a knowing glint in the eye. But if the results are negative or unclear, the reviewers' doubts about the value of your study will be confirmed, and it will almost certainly be rejected. Months of work are at stake. With a *Nature* paper under your belt that early-career fellowship is suddenly within reach. With it would come independence and job security, not to mention prestige and a dose of jealous admiration from your peers.

You run the analysis. The test returns $p = .08$. Damn, so close, but for *Nature* an inch is as good as a mile. What to do? You run through the potential solutions. Collect more data to push the p value below the all-important .05? This is standard practice among your colleagues, but collecting more data is time consuming (you are already approaching the deadline the journal editor gave you), and, in any case, there is no guarantee that the p value would go down rather than up. Also, you're sure you read once that collecting data until $p<.05$ is frowned on by statisticians. You remember asking your supervisor about this once, but the reply was just a snort and a roll of the eyes. No one really cares what the stats geeks say.

Another possibility beckons. Perhaps you can analyze the data differently. Yes, change the rule for excluding outliers from 3 standard deviations to 2. Tweaking the outlier exclusion is legitimate isn't it? It would be harder to justify why you used a different exclusion rule in this experiment from the previous one, but stringency looks good to reviewers, and there's plenty of precedent in the literature for 2. You rerun the analysis. $p = .067$. Close but still no cigar. In annoyance you try 2.5. Then 2.25. Then 1.75. Then 4. Nothing works. The p value is a rock, unyielding.

Another brainwave. Try a different kind of analysis altogether. Yes, that would still be justifiable at a stretch. A few clicks. $p = .057$. Tantalizingly closer, but still not close enough to even be able to round down to the notorious white-lie of "$p = .05$." Something big must be holding you back. You study the individual data. Ten participants. Five show the effect you need

to see. Three don't show much either way. Two show opposite effects, one of which is substantially against the hypothesis. Yes, that's the problem, that one subject with the huge opposite effect. Something must have gone wrong. Perhaps they performed the task incorrectly? Maybe they figured out the hypothesis or didn't follow the instructions? In any case, they're a Bad Subject. You wonder what would happen if you just dropped that subject from the sample—you're pretty sure you could find a reason if needed. Now $p = .052$. Achingly close. Ok, so what if you *flipped* the results for that subject—turned condition A into condition B and vice versa. Just this one subject, just to see what it *would* have looked like if the results were as expected. Now $p = .001$. Magic. Not just borderline significant, but indisputably significant. A scientist-who-knows-what-they're-doing level of significance.

Reality returns with a thud. You changed the data. You touched the untouchable. You click undo. Back to a drab p value that won't budge, and even after applying every questionable (but legal) practice you know, it will never look as impressive as that $p = .001$. No *Nature* paper, then.

But it's late and nobody is around. Nobody else has seen this data. Nobody else has as much riding on this experiment as you do—whether you sink or swim is entirely in your own hands. Who would ever know if you just flipped the raw data for this one subject in this one condition? Sure, it's not The Truth, but your four original experiments were as honest as anyone's, and they all suggest that this experiment really *should* have worked. In any case, the concept of truth is subjective, right? If you alter this data now you'll be able to push the rest of your experiments into a high-profile journal. Then you'll have a shot at securing that fellowship, and you can make a major contribution to science. You would surely be a fool not to take this opportunity. You've never altered data before, and you would promise yourself never to do it again. You've worked so hard and deserve this win. Nobody would ever know. You're already starting to forgive yourself. *Nobody will ever know.*

How many scientists have faced such a dilemma and yielded to the temptation to fabricate results? How many professors achieved their status because of fraudulent behavior earlier in their careers? How many scientists escalate their behavior from the gray area of questionable practices to partial or entire fabrication of data sets? Unfortunately, recent cases suggest that fraud is not as rare in psychology as we would hope. Falsifying data offers a low-risk, high-reward career strategy for scientists who, for whatever

reason, lose their moral compass and sense of purpose. Like many other sciences, psychology is based on an honesty system that is poorly equipped to prevent or detect fraud. Worst of all, when exposed, fraudulent researchers and their institutions are often minimally accountable while whistleblowers face vilification. For turning a blind eye toward fraud and its consequences, psychology perpetrates our fifth major transgression: the sin of corruptibility.

The Anatomy of Fraud

Diederik Stapel liked to win. From a young age, so he writes, he was competitive and enjoyed intellectual challenges such as philosophy, politics, literature, and drama.[1] After failing in the theater, he pursued psychological science and found that it provided an arena that attracted prestige and admiration. Stapel claims to have begun his career at the University of Amsterdam honestly,[2] attempting (via his doctoral work) to stitch together an understanding of various social psychological phenomena that were held together by vague theories built on weak evidence:

> I did a lot of experiments, but not all of them worked. Sometimes I didn't quite get the result I'd expected, and sometimes I got nothing at all. When the latter happens, the best thing to do is to cut your losses and walk away; it didn't work, never mind, on to the next idea. But when I really believed in something—when I really wanted what I'd thought up, or dredged up from the bowels of the literature, to be true—I found it hard to give up, and tried one more time. If it seemed logical, it must be true.[3]

Even at this early career stage it is clear that Stapel was far from the ideal model of a scientist. Like many psychologists he wanted positive results, and moreover he wanted the positive results he expected. He struggled (as any researcher would) to achieve such aims within a culture where low statistical power shrouds most results in uncertainty; where publication bias buries negative results or necessitates *p*-hacking to convert negatives into positives; where failed replications or inconsistencies are nearly always explained by differences in methodology; and where most published findings are regarded (falsely) as true positives. As we've already seen, the out-

come of such biased and weak empiricism is complexity and chaos—a world in which experimenting decomposes into an art. Stapel was keenly aware of this problem.

> Everything had a lot of critical dependencies and the final result could depend on thousands of little things. It was a science of "on the one hand, on the other hand" models and theories, which were successively refined to specify in what situations, under what conditions, at what moment, using what measurements certain effects could be detected. Nothing was always true; everything was conditional on something, and there were always exceptions. For every "rule," it turned out that there was some kind of qualification, some "but" or "perhaps" or "sometimes it works, sometimes it doesn't."[4]

Such a climate makes it difficult for accounts to be falsified—as Chris Ferguson and Moritz Heene have pointed out, the literature quickly becomes a "graveyard of undead theories" that can neither be confirmed nor refuted.[5] Stapel writes that, in his early (more honest) years, some of his peers regarded him as a researcher who formulated overly complex theoretical explanations. In a vain attempt to make sense of the chaos he saw before him, he layered his theories on top of verbose summaries of the literature while his colleagues seemed able to synthesize simpler and more elegant accounts of their data.

> I regularly heard, via the rumor mill, that a senior Dutch academic regularly mocked in his classes what he called my "pinball psychology," because it needed a ridiculous amount of crazy features to keep the ball rolling. A famous American psychologist—one of my heroes—called me a "grinder," by which he meant that I put all the relevant, reliable, robust effects in a pot along with all the edge cases and exceptions, and ground them to a featureless, incomprehensible pulp. He considered that my theories and models were highly sophisticated, but also too complicated to understand. . . . My stories weren't elegant enough, and my results weren't exciting enough, to get published in the top journals. Even my best efforts were seen as mediocre. Not impressive enough, not interesting enough, not innovative, much too complicated. I wasn't good enough.[6]

Faced with his own inadequacy, Stapel appeared to have three choices. The first and most challenging approach would have been to defend the complexity of his theories as the most rational attempt to understand the available evidence. He could have taken aim at the weak evidential standards in social psychology, conducted larger and more conclusive studies (at the cost of achieving lower output), and even attempted to drive improvements in robustness and transparency. Alternatively, he could swallow the insults and carve out a respectable academic career within the boundaries of the status quo—an approach that would still be a struggle given the "publish or perish" research culture and his own apparent limitations as a researcher. This would have placed him on par with many psychologists, fashioning stable but unremarkable careers by engaging in questionable practices to generate publishable stories from their data. Finally—his chosen option—he could immerse himself in the game he believed he needed to win, offering a performance that from the outside appeared rigorous and brilliant yet, in reality, was built on lies. And so Stapel seemed to embrace a kind of scientific nihilism, abandoning the premise that any truth at all can be gleaned from psychological science, and instead focusing his efforts on the personal reward of being published in top journals, pulling in grant funds, and attracting applause. In Stapel's world, studies often became tricks, sometimes ingenious ones based on a careful reading of the existing literature, but nonetheless acts of deception rather than serious attempts to understand the mind and behavior.

> After years of balancing on the outer limits, the grey became darker and darker until it was black, and I fell off the edge into the abyss. I'd been having trouble with my experiments for some time. Even with my various "grey" methods for "improving" the data, I wasn't able to get the results the way I wanted them. I couldn't resist the temptation to go a step further. I wanted it so badly. I wanted to belong, to be part of the action, to score. I really, really wanted to be really, really good. I wanted to be published in the best journals and speak in the largest room at conferences. I wanted people to hang on my every word as I headed for coffee or lunch after delivering a lecture.[7]

Fabricating data wasn't the first port of call for Stapel. He admits that before he reached that stage, he would routinely engage in softer acts of

misrepresentation, deliberately exploiting various forms of *p*-hacking to produce publishable outcomes—in his words, "techniques that could put a healthy looking shine on otherwise mediocre results."[8] He would drop or combine dependent measures, drop entire groups that failed to show effects, and creatively exclude outliers whose results opposed the study hypothesis.

> I knew that this was at the outer limits of what was acceptable, or even beyond those limits in some cases, but . . . I wanted the results so badly. Plus, it worked in an earlier study, and the results were really almost, almost right, and it must be true because it's so logical, and it was such a great idea, and I surely wasn't the only person doing this sort of thing. What I did wasn't whiter than white, but it wasn't completely black either. It was grey, and it was what everyone did. How else did all those other researchers get all those great results?[9]

But questionable practices alone were not enough to satisfy Stapel's ambition, at which point he took the extra step into data fabrication. Following a thorough analysis of the literature, he would design whole (often clever) studies and then, along with them, manufacture entire data sets, thus eliminating the risk of the data failing to confirm his hypothesis. Stapel writes about the first time he fabricated the results of an experiment within his office at the University of Groningen.

> I looked at the array of data and made a few mouse clicks to tell the computer to run the statistical analyses. When I saw the results, the world had become logical again. I saw what I'd imagined. I felt relieved, but my heart was heavy. This was great, but at the same time it was very wrong.[10]

Immediately he began to achieve his goals—the coming years would see keynote addresses at conferences, high-impact papers in prestigious journals such as *Science*, further academic promotion, professional prizes, and a reputation as an academic rock star who knew how to make experiments work.

For Stapel, insecure and craving approval within a discipline where the pursuit of truth was (in his mind) pointless, fraud was the rational choice—a high-reward strategy with low risk of detection that achieved his ambitions and attracted little suspicion. Fabricating data brought order to chaos and

enhanced his career. In a twist of irony, Stapel would advance to senior positions at the University of Tilburg that included teaching of research ethics and proposing a new management model that aimed to help staff escape the "publish or perish" culture.[11] Along the way he would see other researchers successfully replicate his fake results, fueling the delusion that "[w]hat seemed logical was true, once I'd faked it."[12] For more than a decade, Stapel successfully ring-fenced his fraud and so lived a double life: by day a charismatic, high-flying dean; by night an anxious fraudster manufacturing data sets at the kitchen table.

In August 2011, having invented data for over 50 publications, Stapel finally achieved the fame he long desired—but not in the way he anticipated. In an act that curtailed their own careers, Stapel's fraud was exposed by a group of three junior researchers—including at least one of his own PhD students.[13] Following a damning interim report of the case published by Tilburg University,[14] Stapel quickly admitted his guilt. "I have failed as a scientist, as a researcher," he said. "I have adjusted research data and faked research. Not just once, but many several [*sic*] times, and not just briefly, but over a long period of time. I am ashamed of this and I am deeply sorry."[15] In a statement within the interim report, however, he denied committing fraud for personal gain at the expense of the careers of his students and staff. "I must emphasize that the errors I have made were not motivated by self-interest. I do not identify with the picture that has been sketched of a man who has attempted to use young researchers for his own gain. I have made mistakes, but I was and remain sincerely committed to the field of social psychology, young researchers, and other colleagues." Stapel was dismissed by Tilburg University, pursued intensely by the media, and successfully prosecuted (albeit on minor charges) in the Netherlands; he later relinquished his PhD from the University of Amsterdam.

The scale of Stapel's misconduct was unprecedented in psychology and one of the largest discovered in science—never before had such a senior and well-known psychologist been exposed as such a prolific fraudster. At the time of writing, the number of retracted papers involving Stapel has reached 58, making it the third highest worldwide in any area of science.[16] Several of Stapel's PhD students were found to have unwittingly included fabricated data in their PhDs, and although the official investigation concluded that they could retain their degrees, many of their peer-reviewed publications were retracted. Most of his students either left academia or took up posi-

tions at obscure institutions. Stapel, however, continued his role in the education system. In 2014 he taught social philosophy at Fontys Academy for Creative Industries, a higher education institution in the Netherlands. During this time, doubts about his apparent redemption were raised when he was accused of defending himself under a fake identity, a questionable Internet practice known as sock-puppeting. In response to concerns about his returning to teaching, Stapel allegedly masqueraded as a user named "Paul," writing on Retraction Watch:

> Perhaps, God forbid, Stapel is able to teach his students valuable lessons and insights no one else is willing to teach them for a 2-hour-a-week temporary, adjunct position that probably doesn't pay much and perhaps doesn't pay at all. The man is a failure, yes, but he is one of the few people out there who admitted to his fraud, who helped the investigation into his fraud. . . . Nowhere it is written that failures cannot be great teachers. Perhaps he points his students to other frauds, failures, and ridiculous mistakes in psychological science we do not know of yet. That would be cool (and not unlikely). Is it possible? Is it possible that Stapel has something interesting to say, to teach, to comment on?[17]

Notwithstanding his alleged sock-puppetry, Stapel found support in some unlikely corners. Psychologist Jelte Wicherts argued that Stapel had already paid the price for his misconduct and had the right to pursue a teaching career. "He lost his scientific career," wrote Wicherts on Retraction Watch. "He relinquished his PhD title. . . . As far as I know, he did not threaten to sue anyone. He did 120 hours of community service. As long as he doesn't do unsupervised research again, I see no problems with him teaching a course at Fontys."[18]

The final report into Stapel concluded with a number of scathing criticisms of academic life, condemning Stapel's nearly untouchable status within Tilburg University, the failure of senior academics to take two prior allegations of his fraud seriously, the barriers and dangers faced by the whistle-blowers, the prevalence of questionable research practices that obscure fraud, and the failure of scientific criticism to detect his fraud sooner.[19] Among a long list of recommendations to improve transparency, the committee called for an inquiry into research standards in social psychology. At the time writing, many of these recommendations have yet to be imple-

mented—a lethargic albeit unsurprising response. The extreme nature of Stapel's fraud has allowed him to be dismissed by the establishment as a one-in-a-million aberration rather than the predictable by-product of a culture that, in Stapel's own words, rewards "dig and deliver" over "dig and discover."[20]

For many psychologists, the instinctive reaction this case is to believe that fraud is rare, and that Stapel not only manipulated his data but also played the academic system in ways that few other scientists would contemplate. The available evidence, however, suggests that this belief, while comforting, is naive. The scale of Stapel's misconduct was among the greatest known in science, but fraud is by no means as rare or insignificant as many assume. A 2009 meta-analysis by Daniele Fanelli concluded that approximately 2 percent of scientists (including a range of fields other than psychology) admit to engaging in some form of data falsification.[21] In 2012 Leslie John and colleagues further estimated that as many as 40 percent of psychologists have falsified data on at least one occasion.[22] If this statistic is reliable then psychology is a discipline in which there are almost certainly a substantial number of undetected fraudsters currently in operation, perhaps on the scale of Stapel or greater. The question then becomes not *if* Stapel is the tip of the iceberg but *why*, and what if anything can we do about it?

The Thin Gray Line

There is no question that the practices of Diederik Stapel were among the most extreme acts of academic misconduct known, but where should we draw the line between fraud and the gray area of legal but questionable practices that form so-called cultural problems in science?[23] To illustrate this dilemma, consider the following scenarios. Which of the following do you consider to be scientific fraud?

1) A scientist collects 100 observations in an experiment and discards the 80 that run counter to the preferred hypothesis.
2) A scientist runs ten experiments then selectively writes up the two that produced positive results that agree with the preferred hypothesis.

Most would agree that the first scenario is fraudulent. Many (perhaps most) of us would also view the second scenario as fraud, or at least a degree of misconduct. Mathematically, the two scenarios are the same: whether the researcher discards 80 percent of their data or 80 percent of their experiments, they will end up in the same place—with biased evidence. Morally, perhaps they are also on par. After all, how is selectively reporting an experiment based on a desirable outcome any less dishonest than selecting data *within* an experiment for the same reason? But now consider a third scenario:

3) A journal reviews ten papers on the same topic and selectively publishes the two that reported statistically significant effects.

Things just got complicated. In a breath we've ascended from individual dishonesty to the "cultural" problem of publication bias. We can now assert a diminished individual responsibility, forgetting that the incentive structure in scenario 3 ("We publish only positive results") drives scenario 2 ("I'll write up only experiments that produce positive results"), which in turn encourages scenario 1 ("I need this experiment to show positive results"). Unless we define scientific fraud narrowly as (only) the kind of extreme fabrication perpetrated by Stapel, it is difficult, if not impossible, to distinguish the blackest of fraudulent acts from a continuum of dishonest practices at the individual and group level. What about analyzing a dataset 100 different ways and reporting only the analysis that "worked"? Or trawling through methods of excluding outliers to push a p value below .05? Are these acts fraudulent? If not, are we excusing such practices on rational grounds or simply because they are so common? What would your next-door neighbor think?

Stapel himself admits that his journey toward data falsification was gradual, beginning within the gray zone of questionable but common practices in scenario #2 (discussed in chapters 1–3) before escalating to scenario #1 and beyond, all prompted by publication bias in scenario #3. This suggests that deliberate exploitation of questionable practices may provide a gateway to fraud, which in turn raises the tricky question of whether conscious exploitation of questionable practices should also be treated as fraudulent. The case of Dirk Smeesters, introduced in chapter 4, provides an intriguing case in which these questionable practices themselves were judged as misconduct, leading to article retractions and Smeesters's resignation.

We can recall that Smeesters was exposed by Uri Simonsohn after his results appeared too good to be true. In the official report prepared by Erasmus University in 2012, Smeesters confessed to selectively removing data that didn't fit his hypothesis:

> Smeesters admitted that he "massaged" data in three articles . . . to achieve [statistical] significance and that he did not save the raw data adequately, and they were lost as a result. The massaging of data was done using the "blue dot technique" in which subjects who have not read the instructions properly are identified. Massaging takes place by first analysing all of the subjects; if the outcome is not strong enough then the analysis is repeated omitting the subjects that did not read the instructions properly. Smeesters said that data massaging using the blue dot technique is common, and that he [is] therefore part of a culture. He agreed with the committee that this culture is open to discussion and that this does not detract from the fact that he applied this technique without stating so in the papers.[24]

One particularly notable aspect of this case was that the report did not uncover sufficient evidence to conclude that Smeesters fabricated data. The most it could conclude was that he deliberately exploited practices that the majority of psychologists admit to using, and that he also kept bad records. Was this a sufficient reason to prompt his resignation, which was received by the university even before the investigation had concluded? Psychologist Rolf Zwaan from Erasmus University chaired the investigation and believes it was. "He was right to resign," Zwaan told me. "I suspect he feared that further scrutiny would reveal more problems. I don't think he was redeemable."[25] At the same time, Zwaan acknowledges that they had insufficient evidence to demonstrate that Smeesters was falsifying his results in a similar way to Stapel.[26] "We didn't think we could prove that he had fabricated data," he explained. "I don't want to speak for the others [on the committee] but I certainly suspected he did. Our solution was to say: the results are too good to be true and since the raw data are missing, we think that these papers ought to be retracted." At the time of writing, seven of Smeesters's articles in peer-reviewed journals have either been retracted or earmarked for retraction.[27]

Smeesters claimed that the practices he used were commonplace, an allegation that the 2012 survey of US psychologists by Leslie John and col-

leagues would later confirm. Zwaan agrees. "It seems plausible to me," he said. "I have to admit that before being on the committee I was almost completely unfamiliar with his area of research. . . . When I started reading more in that area, I was dumbstruck. During his second interview, Smeesters dropped a stack of papers on the table and said: these people are doing exactly the same thing I'm doing." Responding in a later interview with a journalist, Smeesters claimed that his behavior was motivated by publication bias. "Journals almost exclusively publish studies with statistically significant results, which can indeed lead to massaging your data." He also denied that his resignation was linked to the misconduct case, claiming instead that it was due to "burnout."[28] "I definitely will not miss academia," he said. "Frankly, I am relieved that I am no longer part of it. . . . If I look back now, I took the job because I wanted to do research. But in retrospect my happiest moments were when I returned from a lecture that went well, and the students had treated me to an applause."[29] On that point, at least, Smeesters and Stapel shared one characteristic—the desire to please their audiences.

The Smeesters case sets an interesting precedent in which deliberately engaging in questionable practices—routine procedures that many psychologists admit (anonymously) to exploiting—led to a conclusion of research misconduct and resignation by a tenured professor. However, this standard is not always applied consistently. Months later, the case of German psychologist Jens Förster would show that even when researchers admit to selectively reporting results that support their hypothesis, dismissal can be avoided through strenuous denial of fraud.

The Förster saga began in 2012 when an anonymous whistle-blower noticed that the results in three of Förster's publications looked too good to be true. Over 40 experiments, the papers in question examined the effect of different kinds of perceptual training (termed "induction") on so-called creative and analytic thinking. Each paper reported that training people to focus on perceiving objects "locally" (that is, on a fine-grained scale) induced more critical, analytic thinking, while training them to focus on objects "globally" (on a larger scale) induced more creative, open-minded thinking. For example, upon viewing a large letter H that was composed of many small Fs, responding to the H trained people to think globally. The whistle-blower noticed that the results in the experiments varied too symmetrically around a neutral control condition, leading to patterns that

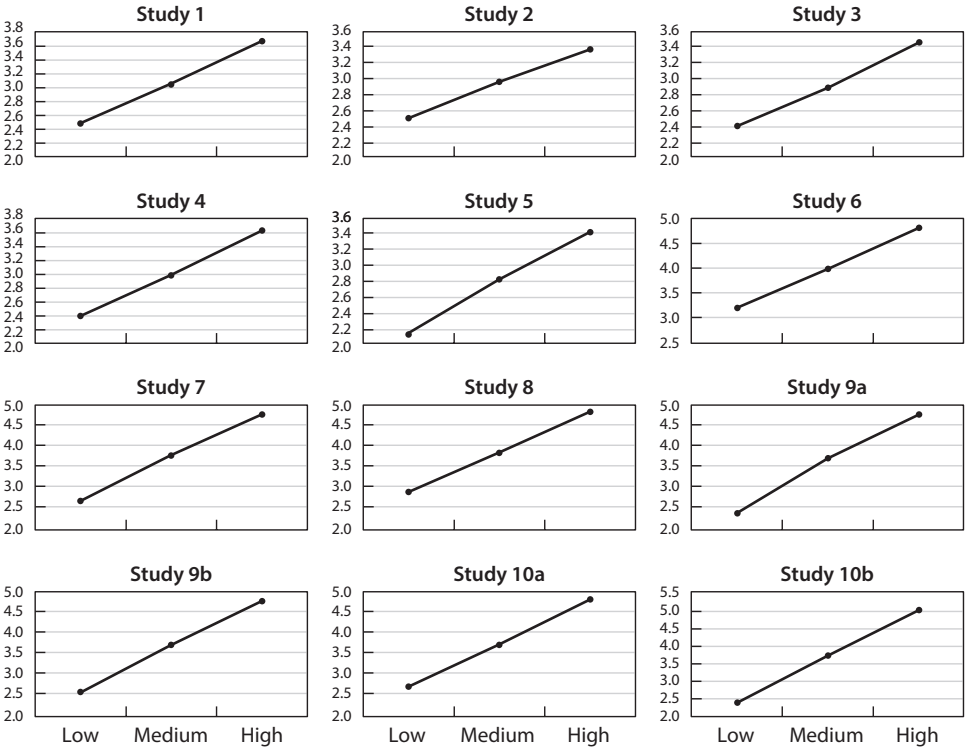


FIGURE 5.1. The 12 studies from Förster and Denzler (2012), showing unusually linear patterns of effects between the local ("low"), control ("medium"), and global ("high") conditions. The paper has since been retracted on the grounds that the data must have been manipulated. Reprinted from the University of Amsterdam internal report.

seemed too linear to be the result of random sampling (see figure 5.1 for an example from one of Förster's papers). A statistical analysis suggested that the probability of finding such perfectly linear (or more linear) results across all three papers was just 1 in 508 trillion[30]—which, put in perspective, is about 50,000 times less likely than two unrelated people having identical forensic DNA profiles or equivalent to winning a national lottery several times in a row.[31] The initial complaint by the whistle-blower also pointed out various other unusual characteristics of the experiments, including implausibly large effect sizes, a gender distribution that didn't match the demographics of the recruited undergraduate population, the complete (and implausible) absence of a single participant drop-out across more than 2,000 participants, and the fact that Förster claimed to have since "dumped" the

raw data. Following an internal investigation, an executive board at the University of Amsterdam recommended that Expressions of Concern be placed alongside the three journal articles. However, according to the whistle-blower, the board fell short of rendering a verdict of misconduct on the grounds that, because the raw data had been destroyed, there was insufficient evidence to conclude that the results were due to fraud as opposed to questionable research practices.[32] Note the contrasting judgments of Tilburg University in the Smeesters case and the University of Amsterdam in the Förster case: while Tilburg University considered deliberate exploitation of questionable practices to be misconduct in the Smeesters case (worthy of retraction and leading to resignation), the University of Amsterdam judged that comparable practices by Förster fell short of misconduct.

Förster's whistle-blower disputed the conclusion of the University of Amsterdam and escalated the case to the Dutch National Board of Research Integrity (LOWI). On condition of anonymity, he told me why he decided to take matters to a higher level. "I didn't feel [the university's decision] was a very smart move because the data were practically impossible and the raw data were gone, and the only outcome could reasonably be to retract the papers." When the case eventually surfaced publicly in 2014, the LOWI had delivered a more damning verdict than the University of Amsterdam. Following consultation with additional experts, the LOWI report concluded that there *was* sufficient evidence of research misconduct. "According to the LOWI, the conclusion that manipulation of the research data has taken place is unavoidable," it stated. "The LOWI found that the variability in the scores for the control group is so unlikely small that this cannot be explained by sloppy science or QRP [questionable research practices]; intervention must have taken place in the presentation of the results of the experiments."[33] The LOWI implied that the responsibility must lie with Förster, ruling out "the possibility that the many research assistants, who were involved in the data collection in these experiments, could have created such exceptional statistical relations."

Förster strenuously denied engaging in either misconduct or questionable research practices. Writing on his blog in 2014, he accused the LOWI of a "terrible misjudgement" and perpetrating a "witch hunt."

> I do feel like the victim of an incredible witch hunt directed at psychologists after the Stapel-affair. Three years ago, we learned that Diederik

> Stapel had invented data, leading to an incredible hysteria, and understandably, this hysteria was especially strong in the Netherlands. From this point on, everybody looked suspicious to everybody.
>
> To be as clear as possible: I never manipulated data and I never motivated my co workers to manipulate data. My co author of the 2012 paper, Markus Denzler, has nothing to do with the data collection or the data analysis. I had invited him to join the publication because he was involved generally in the project.[34]

How, then, did Förster explain the unlikely linear trends in his data? Without providing any clear answers, he appealed to the possibility that the linearity might be a genuine phenomenon, but that he also couldn't rule out fraud by one of 150 research assistants he claimed helped collect the data. "Let me repeat that I never manipulated data," he said. "However, I can also not exclude the possibility that the data has been manipulated by someone involved in the data collection or data processing."[35] Crucially, he also admitted to engaging in scenario #2 discussed earlier, selectively publishing experiments that supported his hypothesis. "Coworkers analyzed the data, and reported whether the individual studies seemed overall good enough for publication or not," he wrote. "If the data did not confirm the hypothesis, I talked to people in the lab about what needs to be done next, which would typically involve brainstorming about what needs to be changed, implementing the changes, preparing the new study and re-running it."

Could selective publication of experiments that "worked" lead to the extraordinary pattern of results in Förster's three challenged publications? Might they form a subset of a much larger file drawer of unpublished results? Although the University of Amsterdam committee concluded that this, among other questionable practices, was a possibility, the whistleblower disagrees. Even selective publication of 40 experiments among hundreds or thousands would still leave odds of "just one in millions" that the published data is genuine. "On top of that, it would require an extreme amount of effort and an extreme amount of experiments and massive numbers of students." Even if such practices were the underlying explanation, they would still warrant retraction of Förster's articles because the data underlying those papers would be heavily biased. More to the point, it is interesting to note that selectively publishing experiments that supported the desired hypothesis—a practice Förster admitted to using—was deemed

insufficient to demand a resignation when comparable practices at the level of the *data* were enough to see the demise of Smeesters.

Unlike Smeesters, at the time of writing Förster remains in the academic system, holding a position at Ruhr-Universität Bochum in Germany. In 2015 he returned a prestigious €5 million grant from the Alexander von Humbolt Foundation that was awarded during the LOWI investigation, writing on his blog that in the wake of his "conflict" that "the organization of a [€]5-Million-project including 50 co workers is impossible."[36] Förster claims to have had an epiphany while hiking in the mountains. "I will spend the rest of my life on BEING rather than on HAVING," he wrote. "Thus, I will leave the materialistic and soulless production approach in science. And I want to say 'Adieu' to 10 cruel years, in which my life was almost completely determined by others. I am going my own way now." Later that year an independent attempt to replicate Förster and Denzler's retracted 2012 study failed rather spectacularly. Author Marc Dressler concluded: "the failed replication, together with the uncovered discrepancies, are not suitable to relieve Förster from suspicion of fraud."[37]

When Junior Scientists Go Astray

Acts of malfeasance by high-profile professors will always make a splash. Stapel, Smeesters, and Förster were all successful researchers holding academic positions at prestigious institutions, and their fall from grace has been extensively publicized, combed over, even glamorized.[38] But what of lesser-known fraud committed by more junior researchers? Such cases may be less newsworthy, but they reveal how misconduct penetrates all levels of the academic hierarchy, and how it can be a tempting strategy for early-career scientists facing uncertain futures and heavy competition.

Few cases exemplify the moral dangers faced by junior researchers more than the case of PhD student Wouter Braet from the University of Leuven in Belgium. Braet became the subject of a misconduct investigation in 2013 when members of his lab discovered serious errors while reanalyzing the fMRI data from one of his previously published papers. The paper in question, originally published in 2012 in the leading journal *Cerebral Cortex*, was subsequently retracted. According to the retraction notice: "[I]t turned out that the original data analyses by the first author [Braet] included several

operations which are hard to replicate and which do not fit fully with the methods as agreed upon with the co-authors and as described in the paper. Because of this we no longer consider these results trustworthy."[39] The same problem was subsequently identified in a second paper, published in *NeuroImage*, which was also retracted.[40]

At first glance, Braet's case might appear as an honest error, or series of errors, rather than premeditated misconduct. However, according to lab head Hans Op de Beeck (and confirmed by the official University of Leuven investigation), while Braet's actions began as honest exploration of the data, they escalated to fraud when Braet deliberately allowed errors in his analyses to remain uncorrected because they generated attractive results.[41] "He never explicitly made the choice to falsify data, in the sense of getting to work one day and saying 'now I am going to manipulate my data, all the way,'" Op de Beeck speculates. "It was more a case of diverting from the proper procedures here and there, in the beginning in a rather exploratory way and later on more systematically, then willingly keeping the errors after finding out that things only worked that way, and then finally not disclosing this to me or other people in the lab."[42]

Like many young scientists, Braet appeared ambitious and strove to publish as many papers as possible in prominent journals. Op de Beeck believes that the pressure Braet felt to get the "right results," together with his lofty career ambitions, fueled a temptation to cheat that began with subtle bias and later escalated into outright deception. "At the time that he made the worst decisions he was getting in data for a control experiment during the process of revising a paper for a high-impact journal," Op de Beeck recalls. "The revision would have been a no-go if he would have analyzed the data without any manipulation. He really needed this paper for his career. He wanted to get a faculty position—an ambition that I supported—and he was even very selective about which places would be good enough for him, which was an ambition that I was trying to tone down because it was unrealistic."

Braet himself claimed that his lurch into fraud was motivated by the "publish or perish" culture of science. Speaking about his case to the Belgian newspaper *De Standaard*, he said: "The publication pressure weighed heavily on me. My academic career wouldn't go anywhere if I did not write enough papers. This pressure has led to the errors."[43] Op de Beeck believes that the temptation for a junior scientist like Braet to commit fraud can be matched

by pressure on supervisors to cover it up, particularly when the supervisor's own career is uncertain. "Imagine if I had been under the same career pressure as Braet when I found out about what he did," Op de Beeck explained. "Imagine if I didn't have tenure yet but was also on a temporary contract and would have been in the process of applying for jobs here and there. It would have been even harder to make this decision [to report Braet's case] than it already was." Op de Beeck is adamant that avoiding conflicts of interest is essential to preserving research integrity. "The more that important decisions are made by people for whom less is at stake, the better."

The Braet fraud shares many characteristics with other cases of junior misconduct. In the same year as the Braet case it was revealed that Adam Savine, a PhD student with Todd Braver in the Cognitive Control and Psychopathology Laboratory at Washington University, had fabricated results in three published papers and six conference presentations.[44] Savine was found out after Braver noticed discrepancies in Savine's data files shortly before he was due to complete the oral defense of his PhD. Like Braet, Savine swiftly confessed to falsifying his findings. His PhD defense was cancelled, and he was reported by Braver for violation of academic integrity.[45]

Both Braet and Savine left academia, but what impact did their actions have on Op de Beeck and Braver? Op de Beeck says he isn't aware of any negative career consequences. "People that know the case well seem to be convinced that I did what I had to do and even see it as a very positive example of what should be done more often," he explained. "The internal university investigation did not blame me or the lab, and promotion committees or other internal committees do not seem to look at it negatively." Nevertheless, he has changed his lab practices in the wake of Braet to emphasize collective responsibility. "It is now made more clear that data are the property of the lab, not of an individual person. We found out about the earlier fraud by a reanalysis of the data, so that makes it very obvious that data might be checked later. We now have a more systematic back-up and archiving system, and the co-workers know that both I and my research manager have access to these data." He has also overseen a cultural change in his lab in the way experiments are planned and analyzed. "As a lab we tend to be more clear now on what sort of decisions should not be taken by one person in isolation and should be discussed in a group, or at least by the main researcher together with me. Decisions such as the number of subjects

to scan, whether or not a subject should be excluded from the analyses, whether we should rely upon analysis A or B, can be quite difficult to make and are prone to experimenter biases of all kinds. We already discussed such matters before the Braet case, but now we do it more consciously and more systematically."

For Op de Beeck, his experience proved how vulnerable senior investigators can be to dishonest practices under their very noses, committed by their most junior charges. "The case has taught me that everything we do is based upon trust," he explained. "I and principal investigators like me can detect honest mistakes, but we are easy to fool." Nevertheless, while Op de Beeck agrees that greater transparency in research practices could help prevent fraud, for instance in allowing easier reanalysis of existing data, he is concerned that knee-jerk reactions to fraud cases could lead to draconian and ineffective regulation of the scientific community. "Even very tight rules, which would be cumbersome for the many honest scientists, would probably still allow the exceptional dishonest scientist to commit fraud if he or she really wanted to," he said. "I remain convinced that we have to start with trust, and implement practices that help the mostly trustworthy scientists do their job as well and as reproducibly as possible, without frustrating them with bureaucracy."

For Braver, the consequences of the Savine fraud have been more profound. At the time the case was initially publicized in 2013,[46] he announced publicly that he was committing his laboratory to practices that can better detect and deter acts of misconduct.[47] Two years later, I caught up with Braver to find out what reforms he put in place. "The main changes have been to better standardize, centralize and make transparent the documentation and archiving process for each project," he told me. "Another change that we are starting to push hard, and transition toward, is adopting the Open Science Framework and principles for our projects." Braver is enthusiastic about open science practices, and in particular data sharing. "I am strongly encouraging each person in our lab to think about using the Open Science Framework as a repository for raw data and analysis files, so that we can make these public when a paper is published. It is a bit more challenging with neuroimaging and other cognitive neuroscience studies involving large amounts of complex data, but we are looking into various solutions."[48]

Like Op de Beeck, Braver says that he isn't aware of any detrimental effects on his career of having (unknowingly) harbored a fraudulent researcher. However, soon after the Savine case broke, the Society for Neuroscience temporarily suspended his membership on the grounds that, despite being found innocent of any misconduct by the US Office of Research Integrity, he nevertheless shared a collective responsibility for the fraud. The society's ethical policy states that it "reserves the right to impose sanctions and/or to take corrective actions, regardless of whether intentionality is demonstrated during an institutional investigation."[49] The policy does not elaborate the specific ethical justification for censuring innocent PIs [principal investigators] or coauthors—a move that smacks of collective punishment. Furthermore, it is difficult to see how such a policy achieves anything beyond dissuading PIs from transparently exposing fraudulent behavior in their own labs.[50]

The cases of Braet and Savine remind us that acts of fraud are not limited to the most senior professors in science; indeed, the temptation to cheat may be just as powerful, if not more so, among junior researchers facing stiff competition for career progression. Had Braet and Savine managed to avoid scrutiny, it is foreseeable that they would have followed the footsteps of Stapel or Smeesters, so it is ultimately beneficial that they were detected early. At the same time, it is clear that for Braver and Op de Beeck, discovering fraud within their own labs was one of the deepest betrayals that can happen in academic life. In 2013, Braver described his feelings at uncovering Savine's fraud as "shocked, angry, and incredibly disappointed . . . the worst kind of violation of my trust, time investment, and efforts towards Adam's mentoring. Being a graduate supervisor is in some ways akin to parenting, so it was like finding out that your child had just skipped town after robbing you and your neighbor's house."[51] Op de Beeck offers a similar sentiment: "On the day I made the final decision to retract the papers and (the inevitable consequence) to go public, I was so upset that I got my car into an accident in the parking lot. The scratches and dent are still very obvious. Nevertheless, the car still drives—a very appropriate metaphor of what happened to me." The silver lining is that Braver and Op de Beeck's swift and transparent responses left them with no lasting career damage. On the contrary, their experiences led to renewed laboratory practices that not only reduce the likelihood of fraud but also help reduce everyday sources of research bias.

Kate's Story

Fortunately for both Braver and Op de Beeck, their discovery of fraud within their own ranks left their own careers shaken but intact. However, when this power relationship is reversed—and it is the junior scientists that uncover misconduct by their seniors—the consequences are more extreme. As we saw earlier, the three young researchers who exposed Stapel were crushed by his fall, either leaving academia or taking up positions at relatively unknown institutions. In other cases, junior whistle-blowers have faced excommunication while the high-ranking perpetrators they identify escape virtually untouched.

The story of Kate illustrates the risks faced by scientists who are courageous enough to identify and challenge senior misconduct. Kate, whose identity has been changed to safeguard her anonymity, spoke to me about what happened when she became involved in a fraud investigation directed at her line manager, and the profound consequences it had for her life and career. The following is based on Kate's account of what happened.[52]

At the time her story begins, Kate was a postdoctoral researcher working for a PI in psychology, based at a major university. Her work with the PI focused on behavioral changes in people who had suffered brain damage and the first year of a three-year position with the PI had gone smoothly.[53] "He was really nice at first, really wonderful," Kate told me. But problems began to emerge during the second year of the job. At the time, the PI was in the process of conducting a single case study in a patient with an interesting type of brain damage. Even though Kate herself wasn't involved in the experiment, the PI had involved several junior members of the research group, including three assistants who were junior to Kate.

Together, the first assistant and the PI had tested the patient using a customized set of equipment, which needed to be positioned in a certain orientation for the experiment to provide a fair test. After it was finished, however, the assistant realized that the equipment wasn't set up in a way that would have allowed such a test, which meant that the results the PI produced (and which happened to perfectly fit his hypothesis) didn't make sense. In the meantime, as part of writing up a scientific paper based on those results, the PI had separately instructed the second assistant to prepare a visual graphic of the equipment. The second assistant duly prepared an illustration of how the equipment was actually oriented in the experi-

ment (i.e., misaligned) and, upon seeing this, the PI instructed her to instead draw it as it *ought* to have been positioned for the results to be interpretable. When the first and second assistants then pointed out this error to the PI, Kate says that he brushed their concerns aside.

At this point all three assistants decided something fishy was going on. One by one they approached Kate for advice. "They each came to me separately and said he was doing something they felt uncomfortable with," she explained. "I told each of them that they needed to talk to each other to get a full picture of what happened. So they had several meetings together, and then with me again, and it became clear that the PI was indeed being dishonest. There was a lot of anger that this could happen because we had never come across anything like it before. I naively thought that if we simply took it to a higher authority then it would be taken care of. So I advised them to see the head of our department." Kate now doubts whether that was the right decision. "In retrospect, this was stupid" she adds. "With everything that happened, what I should have said was drop it, resign, go away."

When the three assistants approached their department head, he told them that if they explained in detail what had happened he would have to raise it with an even higher authority. "After the meeting, they came back to me and said that he had listened," Kate recalls. "But he had also said, are you sure you want to go ahead with this? He wasn't trying to talk them in or out of anything—he was just giving them the options. They decided to pursue it, and to his credit, he raised it with the head of the university." Together the assistants submitted a report to the head of the university that documented their concerns, and, with Kate's help, they were able to provide additional evidence, gleaned from the data itself, that the equipment had been wrongly oriented. A short time later the PI was informed that he was under official investigation for misconduct.

It was at this point that Kate's relationship with her PI deteriorated sharply, despite the fact that she was not one of the official complainants. "He directed all his anger at me. He accused me of orchestrating a campaign against him, of encouraging the assistants to make the complaint. At the same time, he decided he was going to terminate my employment and give my job to a new employee." Kate is convinced that the PI had kept her at arm's length throughout the experiment so that he could more easily get away with fabricating the results. "I was part of several experiments, but

I'm sure he kept me out of this one because he wouldn't have been able to fudge anything if I had been involved. He only included the more junior people—the ones he could manipulate more easily—and even then he was very careful, only giving each one a limited view of the overall picture. That way he could exert his authority more easily. But he underestimated them, and he never anticipated that they would come to me."

Kate recalls that, while the complaint was being considered, the junior team were also placed under a gag order by the head of the university. "He told them that it was now a legal matter and they must not talk to anyone about it, not even privately. As a half-joke, I used to say that he hasn't specifically told *me* that. But of course I remained silent too because I didn't want to bring the department into disrepute. We knew that if we spoke out, people who had nothing to do with the case would suffer. And we believed that everything would be ok because there was going to be procedural justice. We would keep our mouths shut and in return the university would stop him from doing this again."

Kate's faith in the system was misplaced. When the university eventually reached an official decision about the PI's behavior, it fell short of labeling it misconduct, concluding only that he had engaged in "sloppy research practices." He faced no formal reprimand, and no mention of the investigation or its outcome was published. The university's only requirement was that he repeat the disputed experiment. "We were utterly shocked and appalled," Kate says. "By this time each of the junior workers had left to pursue different paths and careers. This left just him and me, that is until he could get rid of me." With the departure of his junior team, the PI chose to rerun the experiment alone. "Given the evidence we presented, I found it extraordinary that the university administration was happy for him to do this," Kate said. "Nobody else was there to monitor what he said to the patient. Nobody else inspected the data." When the results of the experiment emerged, they conformed perfectly to the hypothesis. The manuscript was swiftly accepted for publication in a well-respected journal, where it remains.

In the final months of her second year, Kate felt under constant attack. "I got a lot of bullying from the PI, a lot of petty behaviour. He wouldn't let me finish experiments. He didn't want me back in the lab. In his mind he had done nothing wrong—he was completely and utterly justified and completely innocent of any wrongdoing. His retribution against me was merely

his way of restoring justice." The PI sacked Kate at the end of her second year, and her department was unable to safeguard her employment. With few labs conducting similar research, and few jobs available, she was faced only with hard choices. "The case had taken a great toll, both emotionally and professionally. My choice was to cut and run and only get one publication, or to try and hang on and get more. So I decided to try to go for as long as possible, but obviously that was making it really hard for me."

With the help of a colleague, she was able to take up a temporary research position at a small nearby university to continue her research. Meanwhile she applied on several occasions for faculty jobs back in her original department and each time was unsuccessful. Eventually, after several attempts, she learned that a new head of department had repeatedly blocked her from being shortlisted. Kate is sanguine about this revelation. "At the time it was hard to accept, but in retrospect I can understand this because as a head you have to try to keep a lid on things. Looking back, I was extremely naïve to apply for jobs back there."

After more than seven years of intermittent short-term contracts—and no career progression—Kate finally steered her career back on track by taking a position at a different university. A few years later she then moved to a larger institution, where she currently holds a tenured position. "It took me ten years to get back to where I was," she explains. "And I was even relatively fortunate compared to how it might have played out, because I published several papers during my time with the PI, which means that that period in my career doesn't stand out as an unproductive hole. And I was lucky to know enough senior researchers who could write job references for me, so I didn't have to rely on the PI for those either."

As for the PI, Kate later learned that he was encouraged by the department to move on, and was nudged quietly toward the door by being demoted from a tenured (permanent) position to a probationary one. Within two years of the investigation concluding he had left the university for a post overseas and to this day remains employed as a senior professor in a major institution. In the years to come Kate came to believe that the PI's behavior was not an isolated incident. While she admits she has no hard proof, she is certain he is a repeat offender. "I have very strong suspicions that he fudged a paper in a major generalist science journal and there was a later paper he was working on with another researcher, who happened to be a

close friend of mine, where the data was strange too. She [the researcher] and the PI had sent off the paper and had it reviewed when she pointed out her concerns. They were told by the journal to re-run the experiment and, sure enough, the PI produced a perfect experiment with identical data. But the researcher herself didn't collect the data. She was closely involved with every part of this project but has no idea who re-ran that experiment—it just popped out of thin air. I advised her to leave it alone, which she did, because her contract was ending."

What would Kate's advice be for junior scientists who suspect fraud by a senior researcher in a power relationship? "I would just say, resign, go away, let it go. If I could go back and do it again, I would have finished my own experiments and then I would have resigned and started looking for different jobs." At the same time she is adamant that senior malfeasance must be reported. "There is a difference between reporting it and following it through. The key is that this person—the junior person—should resign. They must resign." The whistle-blower in the Förster case offers similar sentiments. For him, the process took a heavy personal toll, and he offers little to recommend his choices to others. "In hindsight you would say that if you have really clear evidence of misconduct then go for it, but if you think the data are fake but lack definitive evidence, given the system as it is now it is a waste of time."

Even though methods for exposing fraud have advanced since Kate's experience, she believes psychology has a long way to go in adequately addressing the problem. "There have been improvements in stopping it but there is still too much at stake for everyone. Until whistle-blowers are fully protected and institutions are fully transparent, powerful people will continue to get away with it. If they are going to be stopped then it can only be those above them—their line managers and those in the highest positions in the university—that can do it." She also believes that the best way to save junior scientists from her ordeal may be to eliminate fraudsters earlier, as in the cases of Braet and Savine. "We need to be detecting these people when they're young, before they have too much power and before their exposure as frauds causes so much collateral damage. We need to train senior researchers to weed them out from the pool, from postgraduate levels onwards. And we need to be honest when giving references. This would take a whole generation and a major culture shift."

The Dirty Dozen: How to Get Away with Fraud

If fraudulent research psychologists were to compile a playbook on how to beat the system, how would it read? We can glean at least a dozen "top tips" for the aspiring cheat.

1. Take time to read the literature. It contains the material that will enable you generate clever hypotheses and fake data that are both plausible enough and striking enough to impress journal editors and reviewers.
2. Switch raw data between conditions as needed to fit the pattern you need—for example, if you need the measure for condition A to be significantly larger than condition B, make sure that subjects who show large opposite results are switched. Switching is a preferable method to fabricating because the raw data structure is preserved and will carry fewer signs of mishandling. But don't switch every data set that runs against your hypothesis (see #3)—the results need to be believable.
3. When fabricating or switching, never make the results too perfect. Make them good enough to clearly support your hypothesis with low p values that survive stringent multiple comparison correction (this will help create the illusion that you are a rigorous and trustworthy researcher who understands statistics). Remember that squeaky-clean data screams of manipulation and could attract the attention of "data detectives"; passably good data looks more genuine and will usually go unnoticed.
4. Before using your fabricated data, be sure to test it against the fraud detection tools developed by Uri Simonsohn and others. Remember not to get caught out by basic human fallacies such as the law of small numbers. Check the subconditions that are independent of your hypotheses (e.g., men vs. women or younger vs. older) for strange trends that could expose manipulation—it's easy to forget these in haste.
5. Closely guard access to your data. Never share it with outsiders unless forced to do so. Send data to collaborators only if they specifically ask for it, and even then be sure to attach conditions of use. Regardless of who it is, never send more data than abso-

lutely required—remember that many cases of fraud are discovered by the fraudster's close colleagues or students.

6. Use outright fraud sparingly—remember that the best lies are mostly true. Fabricate only the results you need in order to secure a high-impact paper or grant. Fraudsters get caught when they get greedy or complacent. If an honestly conducted experiment disconfirms a fraudulent one, appeal to a hidden moderator, or better still, bury the study in the file drawer where it belongs.
7. Where possible, avoid fabricating results that are likely to be the target of an exact replication. Replications are fortunately rare so this won't be a common problem, but nevertheless be aware that nonreplications by independent groups can draw negative attention and risk exposing you. If, as is sometimes required, you must fabricate results to secure a high-impact paper, make sure the finding is exactly what everyone would have predicted but shown in a clever way—that way, everybody will believe your results and nobody will bother trying to replicate them.
8. Keep your published method sections nebulous on precise detail and ignore e-mails or respond only vaguely to questions from peers about specific methodological practices. Maintaining a certain degree of technical ambiguity will make it harder for other scientists to attempt exact replications of your work, furnishing you with a convenient cover story in case independent researchers fail to reproduce your fake results. If what they find conflicts with your study, you can always discredit them by pointing to some critical aspect of their method that was different from and inferior to yours, even if it wasn't stated explicitly in your paper. Politely decline any attempts to collaborate with "replicators," especially adversarial collaborations where methods are preregistered. These are extremely dangerous, and you should simply say that you are too busy with your own research program to engage in such projects. On the other hand, be sure to note where someone appears to successfully replicate your fraudulent studies (whether a direct replication or, as is more likely, an indirect one). Provided your findings are sufficiently prominent and in line with what others want to see, confirma-

tion bias by other researchers will ensure that at least some of your fake findings are replicated.

9. Take on a lot of interns and unpaid research assistants, dividing their responsibilities so that no one individual sees sufficient workings behind a single experiment to detect any fakery. Keep as few records as possible about who these people are and when they worked in your lab. That way if your data is ever questioned you can pin the fraud on previous lab members, but at the same time your poor record keeping will protect you from having to name anyone in particular.
10. To create an additional smokescreen, write the occasional paper on research practices and, as your career progresses, be sure to take up positions of influence in earnest domains such as ethics and methodology. Don't spend too much time on these (they are only a cover), but building a reputation in sober topics will help frame you as an unlikely exponent of fraud, which in turn will help dissuade potential whistle-blowers from exposing you. At the same time, be sure to oppose greater transparency in research practices on the grounds that science is working just fine, thank you, and has for hundreds of years. Transparency is a threat to your way of life.
11. If ever accused of misconduct, immediately delete your raw data to impede any investigation and dismiss the competence and motivations of your accusers. Where possible, threaten your institution with legal action should they decide to go public. Bully any junior whistle-blowers into silence, and try to draw the sympathy of your superiors and contemporaries by throwing around terms such as "witch hunt," "envy," and "dark forces." Point out that your work has been repeatedly replicated by many others, but without giving specific examples that could be undermined. By making you and your data as untouchable as possible, there is a better chance your university will try to minimize or bury the allegations. And whatever happens, always deny everything. Confess and you will end up on the academic scrap heap.
12. In moments of doubt when you wonder why you are doing all this, remember that science is nothing more than a competition for status in a field of storytellers. You are doing what the system

> requires of you, and, in any case, you have heard of others who are committing far worse infractions than you. After all, it's not as though you are really hurting anyone. Does it really matter if any of your specific results are true or false? Who honestly cares? All findings are to some extent wrong—as a colleague once wrote, the best we can hope for is to be "interestingly wrong." Science always self-corrects in the long run.

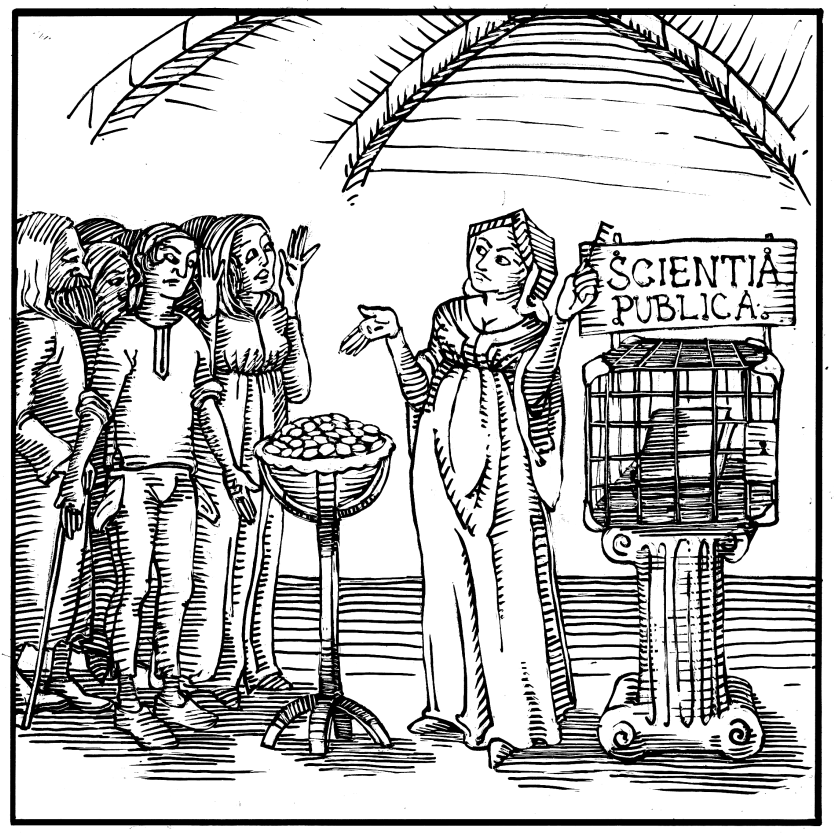
Were psychologists determined to combat fraud then as a community we could do it. But even speaking about the F word, let alone tackling it, is close to taboo. In one hand we brush off fraud as the actions of a few bad apples—a distraction that threatens the public image of science. And in the other hand, we hail the benefits of trust within an honesty system, but without establishing robust systems for detecting fraud and without protecting trustworthy and honest researchers from serial offenders.

How can we bring fraud under control? In chapter 8 we will consider some of the proposed solutions to fraud, including standardized open data, random data audits by an independent authority, and the promise of criminal sanctions for identified perpetrators. We will also return to the ultimate method for weeding out all scientific error, including fraud: direct replication by independent researchers.

CHAPTER 6

The Sin of Internment

Publish means "make public."
—Mike Taylor, 2012



As scientists we celebrate our willingness to share new ideas—after all, science as we know it wouldn't exist without a rich culture of exchange. Yet in chapter 4 we saw how many psychologists reject a key element of this philosophy, refusing to make their data available for scrutiny or reuse by other scientists. Intermingled with this "sin of data hoarding" is an equally worrying, albeit more complex, culture of concealment. Not only are the data supporting research experiments usually unavailable, but the articles that psychologists publish in peer-reviewed academic journals—articles that report the conclusions and implications of their research—are usually published behind subscription paywalls. Publishing behind a paywall means that the public who funded the research in the first place can't read it unless they are prepared to pay again, forking out $30 or more to view single articles, or paying even more to take out personal subscriptions.

Rather than insisting that their research articles are made publicly available, the psychological community has for a long time been content with this system of barrier-based publishing. Most academic psychologists in Western countries cope with access restrictions by taking advantage of subscriptions paid by university libraries (and thus by the public). This allows the academics, but not the public, to read each other's articles, although as we will see not all universities can even afford to furnish their academics with subscriptions. Most importantly, relying on a barrier-based system diminishes the act of publishing from being a means to publicize research outcomes to being an exclusive communication between academics—telegraph lines between the windows of the ivory tower.

This restrictive system is by no means unique to psychology; barrier-based publishing also dominates many other scientific fields. Yet psychology has as much, if not more, reason to oppose such antiquated forms of dissemination. Psychological discoveries generate substantial public interest, are relevant to policy making, and are hugely dependent on public funding. For failing to embrace the culture of open access that the public and non-academic users require, psychology is thus guilty of our sixth major transgression: the sin of internment.

The Basics of Open Access Publishing

The Budapest Open Access Initiative defines open access (OA) publishing as "free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose."[1] This mission seeks to democratize research, making it transparently available to the public whose taxes fund academia. OA publishing also benefits a host of other societal groups, ranging from journalists and science communicators who require timely access to the primary scientific literature, to businesses who can the use latest evidence to drive innovation, to students and professional bodies (such as medical practitioners) needing to learn and apply knowledge, to civil servants seeking to translate scientific discoveries into public policy. When OA is implemented, it generally follows one of two major routes, known as full OA and hybrid OA.

Full OA. The simplest and most accessible OA model is one in which all articles published by a journal are made freely available under a Creative Commons Attribution license (CC-BY) that allows readers to "copy, distribute, display and perform the work and make derivative works based on it only if they give the author or licensor the credits in the manner specified by these."[2] Most fully OA journals can be found to operate in one of two modes: either offering an entirely free service to both authors and readers (generally subsidized by an institution such as a university or library), or by charging authors a fee for publishing through article processing charges (APCs)—the so-called gold OA model. In either case, the cost burden associated with publishing is shifted from readers to authors.[3] Examples of fee-based OA journals include *PLOS ONE* and *PeerJ*, and within psychology specifically, *Archives of Scientific Psychology* and *BMC Psychology*.[4]

Hybrid OA. Under hybrid OA the journal employs a traditional subscription model, but some articles that it publishes will be OA while others appear behind paywalls. There are various paths by which an article in a hybrid journal can become OA. For some outlets, such as the *Journal of Neuroscience* and the *Proceedings of the National Academy of Sciences USA*,

articles are initially published behind a paywall but made freely available to readers after an embargo period (e.g., 12 months), at no additional cost to authors. Alternatively, upon having their article accepted for publication, authors can choose to pay an additional fee to the publisher to make it immediately OA. This gold OA route is analogous to the APC that would be charged for a fully OA journal, and despite the fact that the publisher also collects subscription revenues it typically costs at least 50 percent more. Finally, some publishers allow the author to freely archive their accepted manuscript in a public repository—the so-called green OA route—although most publishers permit archiving of only the accepted postprint version (e.g., produced with word-processing software) rather than the typeset article that appears in the journal itself.

Just how committed are psychology journals to these OA models? Most of the more prominent psychology journals—that is, the journals that psychologists regard as desirable and career advancing—now offer a form of hybrid OA involving both green and gold options. For instance, journals published by the American Psychological Association and the Psychonomic Society offer fee-based gold OA and, in most cases, permit immediate green OA of the author's postprint but not the publisher's typeset version. Other publishers, such as Elsevier, also offer gold OA together with a more restricted green OA in which archiving of the author's postprint within an institutional repository is embargoed for at least 12 months (e.g., the journal *Cortex*).[5] Crucially, very few psychology outlets offer a full OA model, and none of the most prominent and prestigious psychology journals offer a model that is immediately and permanently free for both authors and readers.[6]

Why Do Psychologists Support Barrier-Based Publishing?

Psychology is notable for its tolerance of barrier-based publishing.[7] By comparison, it is standard practice in many physical sciences to archive both preprints (prior to peer review) and accepted postprints (following peer review) in the free repository arXiv.org, which has now been established for 25 years and receives thousands of submissions per month.[8] While physicists still publish in and support peer-reviewed journals (many of which are linked to learned societies), the policies and copyright licenses of these

journals have been shaped around arXiv.org to serve the needs of the community, leading to more transparent and accessible publications. In psychology, however, the practice of archiving preprints and postprints is still rare, even when doing so is fully permitted under the publisher's policy. So what explains the reluctance of psychologists to endorse a more open approach to publishing?

One possible reason why psychologists tend not to archive *pre*prints (despite this being permissible by virtually all publishers) stems from the low reliability of psychological research discussed in chapters 1–3. In psychology it is not unusual for reviewers and journal editors to encourage (or even require) authors to change their hypotheses to fit unexpected results and then, in the interests of generating a clean narrative, to present the revised hypothesis in the introduction of the manuscript as though it was predicted from the outset. Although common, this questionable practice of Hypothesizing After Results are Known (or HARKing; see chapter 1) is frowned on, and since comparing the hypotheses in the preprint with the accepted manuscript would swiftly expose any HARKing undertaken during review, it is perhaps not surprising that few psychologists take the risk of publishing their preprints.

Explaining the relative lack of fully OA psychology journals is more difficult but is probably tied to cultural norms in assessing research quality. In judging the value of published research, psychologists place great weight on journal hierarchy and prestige—values that are monopolized by a small number of prominent, subscription-based outlets. This heuristic for assessing quality isn't necessarily a good thing as the evidence linking journal rank with the merit of individual articles is weak to moderate at best.[9] But here science is bad at being scientific: the actual quality of research plays second fiddle to the *perception* of quality.

In psychology this system stands in the way of OA because very few of the journals that are considered to be high ranking are also fully OA. For instance, in cognitive neuroscience and cognitive psychology, some of the most desirable journals are considered to be *Nature*, *Science*, *Nature Neuroscience*, *Neuron*, *Proceedings of the National Academy of Sciences*, *Current Biology*, *Journal of Neuroscience*, *Psychological Science*, *Cognitive Psychology*, and the *Journal of Experimental Psychology* group. A quick glance at the CV of any top psychologist in this area will probably reveal papers published in this set of journals. And yet, how many of these journals are fully OA? Not

one. At *Nature*, authors don't even have the option to *pay* to make their papers freely accessible.

The "hierarchy of journals" hasn't arisen by chance. Sometimes rank evolves based on history and trust—journals such as the *Journal of Experimental Psychology* group are among the oldest in psychology, having earned a reputation for careful peer review and editing over many decades. Other major outlets ride on the "bling factor," in which prestige is engineered through the selective publication of results that are novel and eye-catching (publication bias). It is not unusual for high-ranking journals to reject 90 percent or more of the submissions received; and the more papers a journal rejects, the more prestigious and desirable it becomes. This isn't necessarily because the rejected papers contained flawed science but because the results are deemed unoriginal, boring, or difficult to hang a story on.

Some proponents of OA have argued for mass boycotts of these journals, calling for scientists (including psychologists) to instead send their work to *PLOS ONE*, *PeerJ*, or one of several emerging (albeit "lower-ranking") OA options.[10] This has a certain appeal. After all, if all psychologists did this en masse then the sin of internment would quickly be resolved. But, of course, no grassroots movement ever happened en masse without a good incentive, and in this case the incentive for psychologists to abandon traditional subscription journals leaves much to be desired: it amounts to sacrificing career opportunities and advancement, including promotions, grants, and research fellowships, for the good of the cause. Publishing in fully OA journals could even lead to scientists losing their jobs. At some UK universities a trend is emerging for the continuing employment of academics to be based, in part, on the quantity of publications they are able to generate in "high-quality journals."[11] And because subscription journals, but not fully OA journals, dominate the list of outlets considered to be "high quality," psychologists face active disincentives to publish their work in more accessible ways.

Concerns about promotion and employability are particularly salient for younger researchers who lack the security of permanent academic jobs. However, while senior psychological scientists have more freedom to publish in less prominent OA journals, the reality is that many continue to reinforce the status quo, not only to maintain their academic influence but to protect the careers of their junior protégés. To illustrate why, suppose you are a professor of psychology supervising a promising PhD student who

you believe could one day become an academic leader. The student has worked hard on a series of ground-breaking experiments and sends a paper to *PLOS Biology*, the only high-ranking journal in your area that is also fully OA. As often happens, though, even for excellent work, the paper is rejected. What now? Do you resubmit the work to a lower-ranking OA journal, hoping that the paper will be recognized for its merits, despite the fact that many scientists at best ignore these journals and at worst ridicule them? Or do you send it to a prestigious journal that sits behind a paywall, where the paper's merits will be boosted by an environment that is both prominent among the desired readership and universally respected? Which is more important, furthering open and transparent science or furthering the career of the scientist whose fate lies in your hands?

This conflict between what's best for science versus what's best for scientists is a real-life dilemma faced every day by senior psychologists who run research groups. Anyone in such a fortunate position has already learned, time and again, that job committees, grant reviewers, and funding agencies are dazzled by articles in prestigious subscription journals. Who, then, is surprised that psychologists at all career levels, from PhD students to professors, refuse to sacrifice their own livelihoods—and those of their protégés—on the altar of OA?

Hybrid OA as Both a Solution and a Problem

How can we change the incentive structure of academia to reward rather than punish psychologists who embrace OA? In the last ten years a number of initiatives have emerged to nudge psychologists toward greater openness. Since 2005, the Wellcome Trust—the UK's largest nongovernment funder of scientific research—has led the way by requiring grantees to publish their academic papers through either full or hybrid OA under a CC-BY license. The trust also expects authors to deposit their manuscript in Europe Pubmed Central, an online database that provides free access to published biomedical research. For the first seven years, this policy was unenforced and, predictably, suffered from low compliance rates with more than half of grantees still publishing their articles behind paywalls.[12] In 2012, the trust clamped down in an effort to boost compliance, with authors who fail to adhere to the policy now losing 10 percent of their grant funding; this

sanction could range from tens to hundreds of thousands of pounds per grant.[13] A review of compliance throughout 2013–14 found that despite some improvement, still only 61 percent of grantees had fulfilled all aspects of the agreement. Remarkably, the trust had also forked out nearly £500,000 in APCs that did not eventuate in publication of the manuscript in Europe Pubmed Central.[14] A later analysis in 2015 was slightly more positive: while the trust had initially withheld funding for noncompliance on 134 occasions, only 20 cases were unredeemable and subjected to sanctions.[15] The trust's head of digital services, Robert Kiley, blames the low compliance on complicated and opaque policies of publishers, particularly in traditional subscription journals where the intricacies of hybrid OA can be most confusing for authors. "These problems are particularly prevalent amongst publishers offering a hybrid OA option, which . . . is also the more expensive way to comply with our OA policy [they were 64 percent more costly]," he wrote. "The Trust remains committed to its OA policy, but ultimately if we are to be successful and get to the point where all Trust funded research is OA, then the friction in the system—which is all too evident in this analysis—has to be resolved."

Recognizing the importance of OA, but also its many barriers and disincentives, in 2012 the UK government commissioned an official enquiry into the future of academic publishing. The report, led by Professor Dame Janet Finch, was unequivocal in its support for the philosophy of OA, declaring at the outset, "The principle that the results of research that has been publicly funded should be freely accessible in the public domain is a compelling one, and fundamentally unanswerable."[16] Notably, however, it fell short of calling for step-change adoption of OA to address this "fundamentally unanswerable" axiom, instead arguing for "a balanced package of measures to be taken to increase access to research publications and to accelerate the transition to open access publishing."[17] Specifically, Finch called for widespread adoption of gold OA where authors pay to publish, rather than green OA (self-archiving), outlining the need for "a clear policy direction . . . towards support for publication in open access or hybrid journals, funded by APCs, as the main vehicle for the publication of research, especially when it is publicly funded."[18] Why the focus on gold OA? Because, according to Finch, "[p]ublishers, whether commercial or not-for-profit, wish to sustain high-quality services, and the revenues that enable them to do so."[19] Indeed, the report clearly supports the commer-

cial needs of the publishing industry in gatekeeping public access to knowledge:

> When funders and institutions began to develop policies to promote open access, especially access via repositories, both commercial and learned society publishers that publish subscription-based journals tended to see them as a threat. Many such publishers saw the prospect of a requirement that articles should be made available through institutional and subject-based repositories, after what was seen as a relatively short embargo period [green OA], as a threat to their revenues and even to the survival of their journals, with the prospect of sales falling as swift, free access became accessible via repositories. Learned societies saw a threat to the publishing income that sustains many of their charitable scholarly and public engagement activities; and also to their income from members who are often attracted by society publications as a membership benefit.[20]

Stevan Harnad, one of the original signatories of the Budapest Open Access Initiative, issued a scathing assessment of the Finch report as a "successful case of lobbying by publishers to protect the interests of publishing at the expense of the interests of research and the public that funds research." He argues that the recommendation to focus on expensive gold OA over free green OA was a costly mistake:

> The Finch Report proposes to . . . push "Green" OA self-archiving (by authors, and Green OA self-archiving mandates by authors' funders and institutions) off the UK policy agenda as inadequate and ineffective and, to boot, likely to destroy both publishing and peer review—and to replace them instead with a vague, slow evolution toward "Gold" OA publishing, at the publishers' pace and price.[21]

Harnad has long advocated green OA, and his call for its adoption makes complete sense within the hybrid model where the publisher also collects subscription fees. However, an exclusive reliance on green OA would also eliminate the growing market of fully OA journals that charge APCs, such as the *PLOS* and *BMC* family. Therefore, at least in the short term, Harnad's proposal to focus on green OA is unsatisfactory because it returns power to the status quo of the traditional subscription publishers. In addition, Harnad's solution would require authors to accede to the embargo restrictions

of green OA imposed by those publishers, meaning that much academic research would be unavailable to the public and nonacademic users until 6–12 months, or more, after it was initially published behind a paywall. In the long term it is possible that community pressure would lower such embargos or eliminate them altogether, which appears to be Harnad's long-term goal. On the other hand, since green OA is effectively subsidized by subscription income, any reduction in subscriptions could lead publishers to apply even greater restrictions on self-archiving in order to protect their revenue streams.

At the same time, the reliance on gold OA advocated in the Finch report is just as unsatisfactory because it applies equally to fully OA journals and hybrid models. Why is gold OA a problem within hybrid models? Recall that with fully OA journals, authors make one payment via an APC to have their article published under a Creative Commons Attribution license (CC-BY) that allows essentially any reuse provided the original work is attributed to the authors. Under the hybrid OA model, however, gold OA—which is on average 60 percent more costly in hybrid journals compared with fully OA journals—adds to existing payments to great an absurd double- or even triple-dipping into the public purse. Consider the farcical nature of the following (completely typical) scenario for a hybrid psychology or neuroscience journal. First a university pays a hefty annual subscription to a publisher, which can amount to tens of thousands per year, per journal, and millions across a "package" of journals. Some of these journals in psychology or neuroscience will then apply additional levies such as submission fees (e.g., at the time of writing, the *Journal of Neuroscience* charges $140), charges for color figures (reaching as high as $200 each), and per-page charges (e.g., $85 per page at the *Journal of Neurophysiology*). And then, on top of these fees, the Finch report would call for the publisher to be paid yet again via a gold OA charge to make the article publicly accessible upon publication—with this fee alone reaching into the thousands per article. The absurdity of this model becomes even more apparent when you consider that in psychology and neuroscience, as in many sciences, the editors and reviewers who coordinate peer review typically do so for little or no financial compensation. What other business in society, other than a journal publisher, is paid three times for a product that is produced almost entirely by someone else?[22] It is easy to see why publishers are content with the Finch report, but it is difficult to see how it offers a sustainable long-term solution or a

justifiable expenditure of public funds. Indeed, it remains a mystery who, other than the publishers themselves, would call the Finch proposal any kind of realistic mechanism for making research openly accessible to the public.

In 2013, Research Councils UK (RCUK)—the central body that coordinates all six major UK research councils, including those that fund psychology research—issued new rules for grantees that cut through some of the controversy and confusion of the Finch report. Since April 2013, publications arising from RCUK grants must now be made OA, either by a green *or* a gold route.[23] By offering authors the choice between self-archiving and APCs, RCUK bypasses the limitation of green- or gold-only solutions advocated by Harnad and Finch, respectively. However, while RCUK's initiative is well balanced, it is also insufficient on its own to open up psychology research. For one thing, it tackles only UK-funded science, which makes up a small fraction of the total global output. And even within the UK, much psychology research is funded directly by universities (rather than through grants) or by alternative funding agencies that do not enforce adherence to OA. In spite of these limitations, there is no doubt that the RCUK model, together with comparable policies launched by the NIH, NWO,[24] the Higher Education Funding Council for England,[25] and European Research Council,[26] are making important headway toward universal OA.

Calling in the Guerrillas

Changes in OA policies are vital, but, as always, top-down reforms can be too slow and bureaucratic to provide immediate benefits for consumers. It is perhaps ironic that the consumers who suffer most frequently from lack of OA tend to be researchers themselves. Even working at a major Western university, I occasionally find that my library will lack a subscription to a particular journal holding a paper of interest. It might be that the journal is relatively obscure, or that the PDF is from an old issue and available for download only under an extended subscription that my library didn't purchase. And so I find myself either contacting the authors for a PDF (a very hit-and-miss approach—in some cases they may no longer be alive) or asking friends at even better funded institutions to send me the article via their more comprehensive library subscriptions. To add insult to injury, it is not

uncommon for researchers at less well funded institutions to lack access to even their *own* published articles. I discovered this in 2003 when, as a young postdoctoral researcher, I had to pay (personally) to download an electronic copy of my very first academic paper, published in the *Journal of Experimental Psychology: Human Perception and Performance*.

The frustration academics experience in accessing the peer-reviewed literature has provoked a remarkable campaign of insurrection against mainstream publishers. In 2011, psychologist Andrea Kuszewski initiated an underground Twitter movement called #icanhazpdf that harnesses crowd sourcing to deliver nearly instant access to the paywalled academic literature.[27] A 2014 analysis estimated that #icanhazpdf is rising in popularity, particularly in the life and social sciences (including psychology), receiving over 3,000 requests per year.[28]

Within the emerging "guerilla open access" movement, some scientists have taken civil disobedience to an even more radical level. In the same year that Kuszewski invented #icanhazpdf, Kazakhstan-based researcher Alexandra Elbakyan launched the website Sci-Hub.[29] Based in St. Petersburg, Sci-Hub provides a search engine and free repository of more than 40 million peer-reviewed academic articles, instantly bypassing publisher's paywalls using a variety of techniques such as obtained (or possibly hacked) academic access credentials. Though the legality of Sci-Hub is in dispute, Elbakyan asserts that it is the corporate publishers who are acting illegally by denying free access to the scientific literature, thereby standing in violation of Article 27 of the United Nations Declaration on Human Rights.[30] Elbakyan's goal with Sci-Hub is a simple one: "to collect all research papers ever published, and make them free," with a particular emphasis on making research accessible to developing countries such as India, Iran, and Indonesia that lack the funds to pay for subscription access.[31]

Acts of subversion can be convenient and powerful catalysts for OA. However, they also run the risk of attracting litigation and even criminal sanctions. For #icanhazpdf, the legality is unclear; since it doesn't involve posting copyrighted material directly into the public domain (only the request for such material is public), it is not obviously illegal or actionable. The legal status of #icanhazpdf remains to be tested in court, but this is unlikely to happen anytime soon. The relatively small user-base of #icanhazpdf presents little threat to the revenues of the major publishers, which can run into billions of dollars per annum.[32] Elbakyan's Sci-Hub, on the

other hand, provides a more clear and present danger to the academic publishing industry. By freely releasing millions of peer-reviewed papers, Sci-Hub provides a route for individuals and institutions to avoid paying journal subscriptions altogether. In an attempt to neutralize this threat, in 2015 Elsevier filed an injunction against Sci-Hub and its affiliates under US copyright law and the US Computer Fraud and Abuse Act, alleging that the defendants engaged in "piracy" and "have caused and continue to cause irreparable injury to Elsevier and its publishing partners (including scholarly societies) for which it publishes certain journals."[33] Four months later, a New York district court granted Elsevier's injunction, after which Sci-Hub and its partner sites were taken down.[34] Within a few weeks, however, the entire Sci-Hub operation reappeared under a different web domain.[35]

In some cases, guerrilla open access tactics have triggered even more than civil litigation, leading as far as criminal charges. In 2011, prominent activist and Harvard researcher Aaron Swartz was arrested by the US Secret Service for downloading millions of articles from an online repository of paywalled journal articles called JSTOR. Swartz had retrieved the articles through his university's JSTOR account, using a laptop hidden inside a wiring closet at the Massachusetts Institute of Technology. After several months of state and federal investigation he was charged with wire fraud, computer fraud, unlawfully obtaining information from a so-called protected computer, and recklessly damaging a protected computer. In 2012, the prosecutor added another nine felony counts to Swartz's indictment, leading him to face up to 50 years in prison and $1 million in fines. Swartz pleaded not guilty but would never see a verdict.[36] On 11 January 2013—two years to the day after his initial arrest—he was found hanged in his Brooklyn apartment.

Counterarguments

Despite the wide range of top-down and grassroots movements to drive forward OA, it nevertheless remains the minority choice in psychology, as in many other sciences. We saw earlier how one of the major barriers in psychology is the low number of prestigious journals that are also fully OA, creating a dilemma for senior scientists striving to promote the careers of

their junior apprentices. This disincentive, combined with the high cost of gold OA in traditional subscription journals, presents a roadblock for many academics. However, to fully understand the reasons for limited support of OA in psychology, we need to look beyond these factors and consider some of the common objections to the OA movement.

Counterargument 1: I have access, so what's the problem?

In the busy life of an academic, the concern that people other than us lack access to the peer-reviewed literature is easy to dismiss. Inertia in OA adoption is fueled, in part, by the fact that the libraries in the richest and most influential universities (mostly, those in major Western countries) are sufficiently armed with subscriptions to make published science available to their own academics. And if influential academics can read the work of other influential academics, doesn't that circumvent the need for OA altogether?

There are at least three problems with this argument. The first is that it relies on the politics of privilege, ignoring the fact that academics at less wealthy universities often lack the access they need.[37] When I was a postdoctoral researcher we once received a visit from a researcher based at a major Russian institution. After two days spent discussing science and what it was like to work in the former USSR, she turned and asked quietly if we could please fill her portable hard disk with PDFs from the major psychology journals. Why? Because her university lacked subscription access to any journals whatsoever. I found it astonishing. How could she function as an academic? How was she able to forge a successful career in psychology? Her solution was a picture of determined industriousness, contacting individual authors one-by-one, by telephone or e-mail, to request copies of their articles. I felt ashamed that such a successful scientist had to ask us to provide access to literature that she had every right to see for herself. I also felt ashamed for assuming, naively, that everyone could see this literature simply because I could. The experience was eye-opening and led me to discover that limited access is a problem experienced to various degrees by many universities, even in Western countries.[38]

The second problem with this stance against OA is that it takes a narrow view on the wider benefits of research for nonacademic groups, including policy makers. Applications of psychology in public policy are many and

varied, ranging from tackling challenges like obesity and climate change through to the design of traffic signs, persuading citizens to vote in elections, and encouraging people to join organ donor registries.[39] In the UK, recognition of psychology in policy making is so prominent that in 2010 the government established a dedicated Behavioural Insights Team, specializing in the application of psychological science to policy.[40] In 2015, a comparable service was established in the United States.[41]

Given the importance of the peer-reviewed literature to policy making, you might assume that the public service would be handsomely equipped with journal subscriptions. Not so. In 2012, Adriana de Palma, a PhD student in ecology at Imperial College London, discovered this for herself when she undertook a short-term placement at the UK Parliamentary Office of Science and Technology (POST).[42] Writing on her personal blog after the placement had finished, de Palma said: "One of the interesting things I learnt while writing my POSTnote [an in-depth scientific briefing for politicians] was the importance of publishing my work open access. The level of access to journals was far lower than I had expected (it was actually shocking)—I ended up using my academic access throughout my placement."[43]

What happens when civil servants lack the means to fall back on academic access credentials, as de Palma could? Steven Hill, head of research policy at the Higher Education Funding Council for England (HEFCE) reports having "zero access to non OA content while working in Defra, RCUK and HEFCE,"[44] an observation confirmed by HEFCE policy researcher Ben Johnson.[45] Johnson even resorted to taking out a personal subscription to gain access—or at least he tried. When even that turned out to be impossible, he issued a public plea for researchers to make their work OA: "If we want more science of science policy, I implore us to deliver the results openly so that those making decisions can actually READ them."[46] On his blog he elaborated:

> If my experiences are the same as others' in my situation, then this is a fundamentally stupid state of affairs for publishing and academia alike. If it is almost impossible for people to subscribe to subscription-based journals—even for honest and willing people like me who (a) are tired of going through the rigmarole of searching out free copies

> of articles, (b) might be willing to pay for a personal subscription, provided it's easy to set up and not too expensive, and (c) are pretty savvy when it comes to using technology and the Web—then what hope can publishers and academics have of their work reaching a wider audience? I despair.[47]

Even putting aside the wider benefits of OA, the argument that "I have access so everything is fine" has a third problem: those with the greatest access can often find that their ability to use subscribed material is encumbered with restrictions. In one recent case, Chris Hartgerink, a PhD student in psychology at Tilburg University, collided with regulatory barriers when attempting to content mine information in academic articles. Hartgerink's research focuses on extracting summary data from published psychology papers in order to detect signs of data fabrication, an approach that requires the automated, bulk downloading and analysis of articles from behind corporate paywalls.[48] Despite the fact that his approach is legal and in accordance with his university's journal subscriptions, Hartgerink reported that Elsevier blocked him from conducting his research:

> Approximately two weeks after I started downloading psychology research papers, Elsevier notified my university that this was a violation of the access contract, that this could be considered stealing of content, and that they wanted it to stop. My librarian explicitly instructed me to stop downloading (which I did immediately), otherwise Elsevier would cut all access to Sciencedirect for my university.[49]

In response to Hartgerink's blog post, Elsevier advised him that his content-mining approach was acceptable only if he used a specific software program supplied by the publisher.[50] Use of this program, however, is conditional on authors agreeing to various restrictions, such as the way in which the extracted data can be published by the researcher, and the capacity for the data to be reused by others. For example, the policy requires that the published content must contain no more than 200 characters of mined text per "snippet," and that it must also be published under a more restrictive CC-BY-NC license that forbids future commercial applications of the work, as compared with a more open CC-BY attribution license.[51] For these reasons, Hartgerink decided that he was unable to use the tool, writing that

Elsevier's policy "harms individual researchers and the impact of research on society."[52] Thus, even within a well-resourced Western university, restrictions on the use of legally subscribed content can fail to deliver the needs of the psychological community.

A broader concern is that, even for the richest universities, university library budgets are under increasing pressure from ever-rising journal subscriptions. In 2012, the Harvard University library issued a provocative and widely publicized memo in which it condemned the growing cost of subscription publishing and called for researchers to widely support cheaper OA journals:

> We write to communicate an untenable situation facing the Harvard Library. Many large journal publishers have made the scholarly communication environment fiscally unsustainable and academically restrictive. This situation is exacerbated by efforts of certain publishers (called "providers") to acquire, bundle, and increase the pricing on journals.
>
> Harvard's annual cost for journals from these providers now approaches $3.75M. . . . Some journals cost as much as $40,000 per year, others in the tens of thousands. Prices for online content from two providers have increased by about 145% over the past six years, which far exceeds not only the consumer price index, but also the higher education and the library price indices. These journals therefore claim an ever-increasing share of our overall collection budget. Even though scholarly output continues to grow and publishing can be expensive, profit margins of 35% and more suggest that the prices we must pay do not solely result from an increasing supply of new articles.[53]

As of 2015, all scholarly articles published at Harvard University are now made fully OA upon publication.[54] Similar steps toward universal OA have been made by the Association of Universities in the Netherlands (VSNU), the centralized body that represents all Dutch universities. Throughout 2014 and 2015, VSNU was publicly vocal in its intention to boycott Elsevier, with the aim to make all publications arising from Dutch universities fully OA. Under the plan outlined by the undersecretary for education, culture and science, Sander Dekker, Dutch universities will cancel all journal subscriptions by 2024.[55]

Counterargument 2: Most of the public don't want OA. Even if they could read the literature they wouldn't understand it, and they might dangerously misunderstand it.

In discussions with colleagues I have found that this is a remarkably common view that stems from an element of truth. As in other sciences, many empirical papers in psychology are highly technical and written exclusively for a specialist audience of like-minded scholars. Theoretical assumptions, background studies, and technical terms are rarely explained in detail because they are (quite rightly) assumed to be common knowledge among the targeted readership. Given that this is the case, it is surely unlikely that basic psychological science, as communicated in specialist journals, would be interesting or intelligible to the general public. Prominent OA critic Daniel Allington has claimed that OA therefore "amounts to no more nor less than the handing of a publicly-funded megaphone to the UK's most privileged scholars. Soon the whole internet will be able to hear them talking to one another."[56]

Although on its face this argument appears sensible, its critical flaw lies in the loose definition of "the public." By lumping together all nonacademics into a uniform mass, we slip into the well-known psychological trap of "outgroup homogeneity" in which individuals not belonging to a narrow ingroup (in this case university academics) are seen as more alike than they really are.[57] In fact, there are many varied applications of specialist knowledge within nonacademic professions that make up a sizable chunk of "the public." The website whoneedsaccess.org provides examples of how OA is important for some of these nonacademic groups—in addition to civil servants and policy makers (as discussed earlier), the list includes translators, research organizations, small businesses, people working with the developing world, doctors and dentists, nurses, teachers, consumer organizations, patients, patient groups, amateur scientists, Wikipedia contributors, bloggers, science communicators, unaffiliated scholars working into their retirement, enterprising schoolchildren, professional nonacademic researchers, independent researchers, publishers, and even artists.[58] Academics who opine that OA isn't necessary because nobody outside their narrow circle cares about their work not only underestimate the value of their own research but also the intelligence and intellectual vigor of groups beyond their immediate horizon.

What of the concern that OA might lead unqualified readers to misunderstand the research, possibly even to the extent that they might put themselves or others in danger? This may be an issue in a limited subset of psychological fields, but even where this is a potential problem, the answer is not for the work to be locked away behind paywalls to "protect" the public but for academics to prepare intelligible lay summaries of their work that dispel likely sources of confusion. There are now many avenues academics can use to engage in productive dialogue with public audiences, for instance by writing articles for outlets such as the *Conversation* and the *Guardian* Science network, by writing on their own personal blogs, or through journalists in the print and broadcast media.

Finally, it is of course true that a large proportion of the public will have no interest in reading any particular peer-reviewed article. However, this provides no escape from the leading axiom of the Finch report: that given the reliance of the research on public funding, who are we to deny public access to the output of our research?

Counterargument 3: OA journals don't meet the same editorial standards as traditional journals because authors pay to publish. The journal therefore has a commercial incentive to accept poor papers.

We don't need to look far in psychology to find this criticism of OA journals—a claim that seems to be particularly prevalent among a certain "old guard" of psychology researchers. One of the most public examples of hostility toward one OA journal in particular was the 2012 *Psychology Today* article by John Bargh, which we discussed in chapter 1. Recall that Bargh was responding to a failure to replicate one of his previous and highly influential "social priming" studies, in which he concluded that showing young participants words associated with elderly people caused them to walk more slowly. In a subsequently deleted blog post, Bargh took aim at *PLOS ONE*—the journal that published the failed replication of his work—declaring that, as an OA journal, it provides substandard peer review:

> PLOS ONE . . . quite obviously does not receive the usual high scientific journal standards of peer-review scrutiny . . . instead, the jour-

> nal follows a "business model" in which authors pay to have their articles published (at a hefty $1,350 per article). . . . If I'd been asked to review it (oddly for an article that purported to fail to replicate one of my past studies, I wasn't) I could have pointed out at that time the technical flaws, though these might not have mattered to PLOS ONE—as a for-profit enterprise, PLOS published 14,000 articles in the year 2011 alone. Fourteen *thousand*. Something tells me they don't turn down many $1,350 checks.[59]

In fact, contrary to Bargh's claims, the paper in question did receive in-depth peer review and was edited by Professor Jan Lauwereyns, an internationally recognized expert in cognitive psychology.[60] Furthermore, *PLOS ONE* selects which manuscripts to publish based not on the authors' ability to pay (which is considered only once manuscripts are accepted, with fee waivers available), but according to structured publication criteria that are clearer than at most traditional psychology journals.[61] Finally, not only is there no evidence to support Bargh's accusation of editorial negligence, his criticism of *PLOS ONE* ignores the fact that many prestigious subscription journals also operate "business models" that charge authors to publish (e.g., through page charges) in addition to charging libraries for subscriptions.

In addition to prejudice against OA journals from senior academics, the OA movement has also come under direct attack from subscription publishers. In 2013, journalist John Bohannon of *Science* magazine unveiled the results of an audacious sting in which he sought to test the quality of peer review in several hundred OA journals.[62] With the help of academic colleagues, he wrote a bogus paper with obvious flaws (and fake results) that should have been readily picked up through peer review. He then submitted the paper to 304 OA journals that charge APCs, including 167 that were listed on the Directory of Open Access Journals (DOAJ), an online database that purports to index "high quality, open access, peer-reviewed journals."[63] The results were disquieting: of 304 journals, over half accepted the bogus paper, often with little or no peer review. Even more worryingly, that number dropped only slightly, to 45 percent, when considering the OA journals listed on the (supposedly selective) DOAJ.

Given the results of Bohannon's sting, aren't scientists (including psychologists) sensible to be suspicious of publishing in OA journals? Wasn't Bargh right after all? Certainly there is reason to be careful about choosing

where to publish—and this is true for both subscription and OA journals—however the sting provides no reason for psychologists to turn away from all OA journals. First, the most prominent OA journals swiftly rejected Bohannon's bogus paper, including *PLOS ONE* and the targeted *Frontiers* journal. In the case of *PLOS ONE*, Bohannon even noted that it "meticulously checked with the fictional authors that [ethics] and other prerequisites of a proper scientific study were met before sending it out for review. *PLOS ONE* rejected the paper 2 weeks later on the basis of its scientific quality." Second, viewed as a research study, the sting also included an elementary flaw—that by selectively targeting OA journals, and not including a control group of subscription journals, it is impossible to know whether OA journals are any more likely than subscription journals to accept bogus papers. Addressing this error on his personal blog, *PLOS* founder Michael Eisen wrote:

> We obviously don't know what subscription journals would have done with this paper, but there is every reason to believe that a large number of them would also have accepted the paper. . . . Like OA journals, a lot of subscription-based journals have businesses based on accepting lots of papers with little regard to their importance or even validity. When Elsevier and other big commercial publishers pitch their "big deal," the main thing they push is the number of papers they have in their collection. And one look at many of their journals shows that they also will accept almost anything.[64]

Eisen also took aim at the selectively chosen subset of OA journal targeted in the sting, which he argued was unrepresentative of OA journals overall: "To suggest . . . that the problem with scientific publishing is that open access enables internet scamming is like saying that the problem with the international finance system is that it enables Nigerian wire transfer scams." He concluded that the "long con" orchestrated by subscription publishers far outweighs the actions of the obviously predatory OA journals who fell for Bohannon's sting: "Not only do [the subscription publishers] traffic in billions rather than thousands of dollars and [deny] the vast majority of people on Earth access to the findings of publicly funded research, the impact and glamour they sell us to make us willing participants in their grift has serious consequences."

Counterargument 4: What's in it for me?

In the competitive world of academic science, researchers naturally seek rewards for engaging in practices that break from the norm. As with data sharing (see chapter 4), there is now substantial evidence that OA publication is associated with increased citation rates. In a key 2005 study, Chawki Hajje, Stevan Harnad, and colleagues from the University of Quebec reported that, across a sample of more than 1.3 million articles published between 1992 and 2003, those that were OA attracted between 25–250 percent more citations than those behind paywalls. In particular, they found that OA articles in psychology were associated with more than double the number of citations.[65] Since then, the SPARC Europe project has catalogued this effect across multiple fields and many published reports, finding that of 70 published investigations, 46 find evidence of an OA citation advantage, with the magnitude ranging from less 5 percent to over 500 percent.[66]

One caveat that must be applied to such studies is that they relied on observational designs in which data from previously published papers were collated and analyzed. Because of their retrospective nature, such designs are unable to show that differences in citations between OA and non-OA papers are caused by the openness of the publishing route as opposed to a myriad of other possible confounding variables, not least of which may be the inherent quality of the research. In order to provide causal evidence that OA boosts citations, we would need to see the results of a controlled trial in which papers were randomly assigned to either an OA or non-OA group, much like a clinical drug trial—and no such interventional study has yet been reported. Nevertheless, the correlational evidence is at least suggestive of a causal effect, so in the quest for professional rewards, the citation advantage of OA provides a good reason for authors to prefer more open publishing practices.[67]

An Open Road

What does the future hold for OA publishing in psychology? Can the field break free, once and for all, from the grip of subscription publishers that can be paid not once, not twice, but three times for a product they didn't

create in the first place? Although we are some way from an ideal future of universal OA, the signs are promising that it will eventually come to pass. In addition to policy-led reforms and subversive grassroots initiatives, we are also seeing novel publishing models emerge in psychology that are creating a new arena for open practices. Recent examples include the new OA journal *Collabra*, which unlike most journals pays reviewers and editors for their work, but cleverly provides them with the choice of either collecting the money directly or paying it forward into an APC waiver fund that can be used by other authors or their own institutions. With wide uptake, this model could be transformative. Meanwhile, *PeerJ* offers a much cheaper gold OA model than other journals by providing membership plans to authors rather than charging APCs per paper. Although neither model or journal is yet regarded as a first choice for most psychologists, the field is positioned for new ideas to take hold.

As we will see in chapter 8, there is an even more radical solution to the sin of internment—one that is not as distant as it might appear. We could abolish journals altogether.

CHAPTER 7

The Sin of Bean Counting

Not everything that counts can be counted, and not everything that can be counted counts.

—William Bruce Cameron, 1963

Science is all about measurement. Whether it's the speed of a neutrino or the reaction time of a person, quantifying natural phenomena is what scientists do. Yet, for all our ingenuity, one question has proven particularly challenging: can science measure itself? And, assuming that the quality of science is something that is in fact quantifiable, can scientists themselves be boiled down to a series of numbers that reflect their past achievements and future potential? In psychology, as in many other sciences, there is a growing push toward weighing up the worth of individual academics and their research contributions based on various "metrics," and to then use those metrics to award jobs and funding. Metrics such as the "impact factor" of the journals in which we publish, the quantity and monetary value of grants we receive, the number of papers we publish, our author position on those papers, and the number of citations we receive, have all become so-called key performance indicators. And because academics, and the bureaucrats who oversee us, view these indicators as synonymous with high-quality science, the measures themselves have become career targets.

There are reasons to be skeptical about reducing the quality of science to numbers. Goodhart's law of economics warns us that *when a measure becomes a target, it ceases to be a good measure*.[1] This is because those of us who stand to win or lose immediately seek ways to game the system in order to achieve the target, sacrificing other important (and unmeasured) goals along the way, and thereby compromising the value of both the measure and our wider mission. On top of this, many such measures hold no true value in the first place.

For basic scientists, the decisive indicator of our empirical achievements is the advancement of theory that accurately predicts and explains future observations. As we saw in chapter 3, this path eventually leads us to refine theories into such razor-sharp explanatory instruments that they become laws. For applied psychological scientists, the goal is more practical, but the logic is the same: the greatest signifier of success is that we changed something in society, whether addressing a problem such as obesity or mental illness, using psychological evidence to develop a successful policy, or building something that in a demonstrable sense "works." Yet in our obsession to quantify quality—to turn a thing from words into numbers—we have allowed these aspirations to become overshadowed by a culture of metrics-

based performance management. In doing so, psychology commits our seventh and final transgression: the sin of bean counting.

Roads to Nowhere

In psychology, the sin of bean counting manifests in several different ways. We reduce the skills and knowledge of individual scientists to their citation metrics and other simplistic surrogates of "impact." We reward academics disproportionately for winning research grants, summing scientific inputs (funding) to outputs (discoveries) rather than weighing them against each other, and in the process favoring more expensive research. And we tally the authorship order of scientists on research papers while clinging to an antiquated system of attribution that muddies their individual contributions. Some of these problems are more severe than others—and none are unique to psychology—but they all stem from the same root cause. To borrow the words of Oscar Wilde, we have become obsessed with the price of everything and the value of nothing.

Impact Factors and Modern-Day Astrology

The first hint of the metrics revolution came in 1955 with the publication of "Citation Indexes for Science," by Eugene Garfield in *Science* magazine. Prior to this there had been no systematic way for scientists to track which academic papers were cited by which, or for counting the number of citations received by individual articles or journals. Garfield's ambitious plan was to bring coherence and efficiency to science by creating a massive referential database of citations. Among his core suggestions was the creation of a citation metric called the *impact factor*—a way of quantifying the influence of a paper on others by counting the number of citations it received. Garfield wrote:

> In effect, the system would provide a complete listing, for the publications covered, of all the original articles that had referred to the article in question. This would clearly be particularly useful in historical research, when one is trying to evaluate the significance of a particular work and its impact on the literature and thinking of the

> period. Such an "impact factor" may be much more indicative than an absolute count of the number of a scientist's publications.[2]

Since then the impact factor has evolved to become one of the most powerful metrics in science—a proxy for the prestige of journals and the caliber of the scientists who publish within them. The modern form of the impact factor is calculated at the level of the journal (rather than article) as the average number of citations in the current year to articles published within the previous two years. So, for example, to calculate the 2014 impact factor for *Nature* magazine we add together the total number of citations in 2014 to *Nature* articles published in 2012 and 2013 (71,677) and divide it by the total number of citable *Nature* articles during the same period (1,729). The answer, 41.5, is among the highest in scientific publishing. Journals are now being routinely ranked by their impact factors, with academics—including psychologists—striving to publish their papers in so-called high impact journals.

In the decades following Garfield's call for science to get organized, he would go on to create the Science Citation Index (and Institute for Scientific Information, now owned by Thomson Reuters), as well as ushering into existence the disciplines of *scientometrics* and *bibliometrics*—divisions of information science that seek to itemize the impact and value of science. Writing in the inaugural issue of *Scientometrics*, a journal that Garfield helped establish in 1978, his coeditor Mihály Beck defended their mission to quantify quality:

> The quantitative evaluation and intercomparison of scientific activity, productivity and progress seems to be sheer nonsense for many, perhaps for the majority of active scientists. This attitude is quite natural, but its source is mainly prejudice, ignorance and/or misunderstanding of the basic ideas of scientometrics. The skeptics are scornful of the hopeless aim of measuring the unmeasurable. The tremendous increase of the scientific production over the last decades has made the emergence of this new field of science both necessary and possible.[3]

Garfield says that he created journal impact factor (JIF) with two purposes in mind: to help libraries decide which journal subscriptions they should purchase, and to help authors decide in which journals they should publish—operating on the (since upheld) assumption that scientists gener-

ally view journals with higher JIFs as being more prestigious and desirable.[4] Garfield's assumption that citation statistics tell us anything about the quality of published work is widely disputed, but not nearly as contentious as the now routine practice of evaluating the quality of individual articles, and individual scientists, by the impact factors of the journals in which they publish.[5]

Why is it a problem to assess the value of individuals based on journal-level citation statistics? The first concern lies in the statistical properties of citation distributions. Most of the citations in a journal are generated by a minority of articles, a phenomenon that Garfield himself has referred to as the "80/20 rule": across science as a whole, around 80 percent of citations are attracted by around 20 percent of papers. This leads to a right-skewed distribution of citations in which most articles receive few citations while a scattered minority receive a lot. The JIF is therefore a poor indicator of central tendency because it represents so few individual samples in the population. Figure 7.1 shows this graphically by plotting the citation distribution for *Open Biology*, a journal published by the Royal Society. Only 1 in 16 articles attracted the same number of citations as the JIF, and many received few to no citations at all. Because of the skewness and spread of citation distributions, a second concern is that JIF only barely, if at all, predicts the citation rates of individual articles within the journal.[6] This means that publishing in a so-called high-impact journal provides no promise that an article will be highly cited, just as publishing in a lesser-known journal is far from condemning it to obscurity.[7] Troublingly, however, JIF does correlate with one metric: the rate of articles retracted owing to fraud or suspected fraud.[8] Rather than referring to outlets such as *Nature* and *Science* as "high impact" journals, it would, if anything, be more accurate to call them "high retraction" journals.

A third problem is that the JIF is easily massaged by editors and publishers. One tactic is for the journal to publish papers that are predicted to be highly cited (such as review articles) toward the beginning of the year, giving those articles more time to accrue citations over the two-year period that feeds into the JIF calculation. For the same reason, it is not unheard of for psychology journals to strategically delay publication of potential "heavy hitters" until the start of the following year. More insidious tactics include journals publishing editorials that excessively cite the journal's own articles, or editors nudging authors of submissions to do the same. Such practices are rightly frowned on but usually fly below the radar of wider scrutiny.

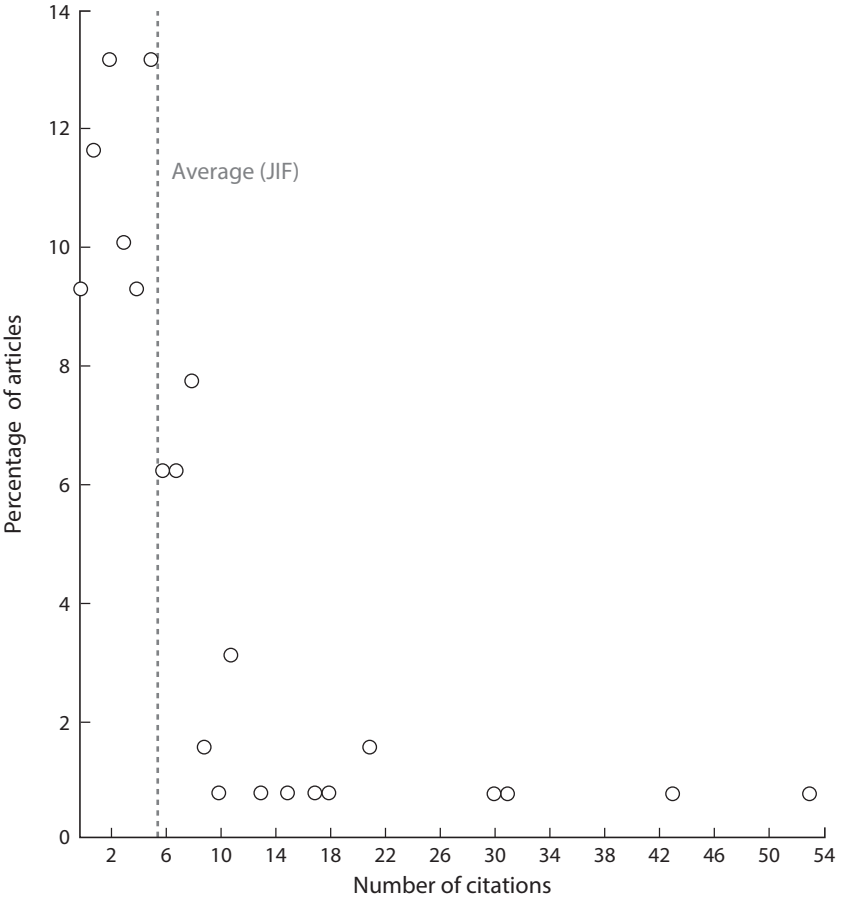


FIGURE 7.1. Citation distributions highlight the meaningless nature of the journal impact factor (JIF). Here we see the distribution of citations used to calculate the 2014 JIF for the Royal Society journal *Open Biology* (the results shown here were extracted from http://rsob.royalsocietypublishing.org/citation-metrics). The chart plots the percentage of articles published throughout 2012–2013 (*y* axis) according to how many citations they received in 2014 (*x* axis), with the JIF calculated as the average of this distribution (vertical dotted line). Note the long rightward tail, or right skew, in which most articles received few citations, with about 9% attracting none at all. *Open Biology* is actually less severely skewed that many scientific journals—here, the top 20% most highly cited papers attracted 56% of the total citations. Even so, the degree of skewness is sufficient to ensure that two-thirds of articles had *fewer* citations than the JIF, and the JIF itself represented only about 6% of individual articles published. As these distributions become broader and more heavily skewed, JIF becomes increasingly meaningless as an indicator of individual article citations.

Even putting aside its statistical flaws and vulnerability to manipulation, the integrity of JIF is undermined by the fact that it is arbitrarily *negotiated* rather than objectively calculated. Journals can lobby the company that determines the JIF (Thomson Reuters) to reduce the number of citable articles in the JIF calculation and thereby inflate the score. In 2006, the editors of *PLOS Medicine* described the remarkable series of face-to-face meetings, phone calls, and e-mails as part of negotiating their 2005 JIF. Depending on which articles were deemed to be citable, their resulting JIF could vary between 3 and 11—a range that would dramatically affect the journal's published JIF ranking. The editors were less than enamored with the opacity and capriciousness of the process, concluding that "science is currently rated by a process that is itself unscientific, subjective, and secretive."[9] In 2013, Björn Brembs, Kate Button, and Marcus Munafò further revealed how the JIF of *Current Biology* jumped by 70 percent, from 7.0 in 2002 to 11.9 in 2003, not because the number of citations increased by a significant amount, but because the number of citable items during the 2001 period was quietly reduced from 528 to 300.[10]

Even if JIFs were indicative of article-level citations (which they are not), would it be safe to assume that the number of citations an article receives reflects the value of its contribution to science? In other words, to what extent is the quality of science synonymous with short-term popularity? To test the idea that reproducibility correlates with JIF—and therefore that article prestige predicts at least one potential measure of study quality—Björn Brembs and colleagues examined the relationship between JIF and statistical power. They found that the statistical power of 650 neuroscience studies was unrelated to the JIF of the outlets in which they were published (see figure 7.2). Not only do such findings call into question the typical methods used to evaluate researchers in academic departments, they also cast doubt on Eugene Garfield's claim over ten years earlier that JIF "is not a perfect tool to measure the quality of articles but there is nothing better."[11] Whether there is any metric better than JIF is unclear, but it is hard to imagine anything worse.

Despite all the evidence that JIF is more-or-less worthless, the psychological community has become ensnared in a groupthink that lends it value. As a junior postdoctoral researcher, I had above my desk a printed list of "top journals" I aimed to publish in, rank ordered by JIF. *Science*, *Nature*, *Nature Neuroscience*, *Neuron*, *PNAS*, *Current Biology*, and so on. I would send

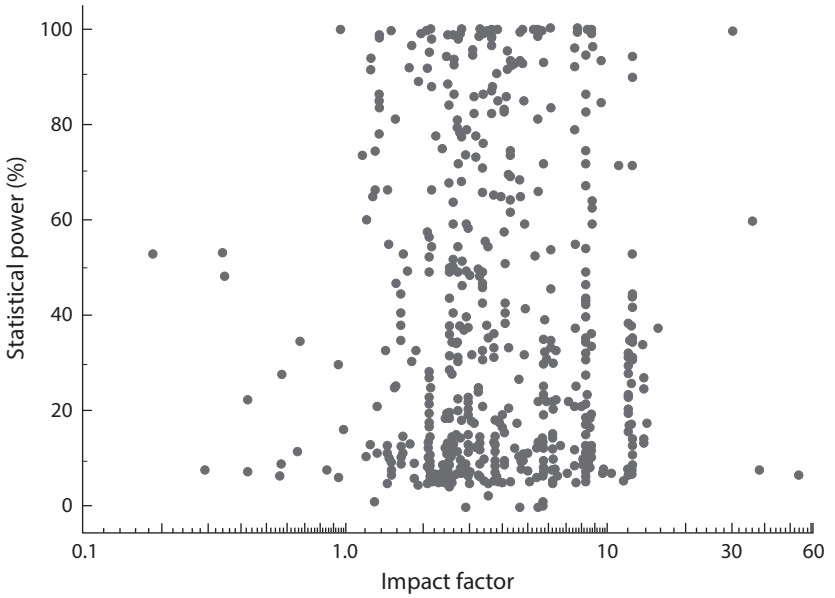


Figure 7.2. Brembs, Button, and Munafò (2013) found no discernible correlation between the statistical power of individual studies in neuroscience and the JIF of the journals in which those studies appeared. These results therefore provide no evidence that journals with higher JIFs publish more reproducible science.

most of my papers to The List, starting near the top and working my way down, one-by-one, until the paper was either accepted in a "high-impact" outlet or I had to settle for publishing in a specialist journal. In retrospect it was a colossal time sink, and not only for me. Consider the waste I produced in the review system—the hours of reviewers' time I consumed by requiring other scientists to repeatedly read and then reject my submissions, not (in most cases) because my studies were scientifically flawed but because the research question or results were not important or definitive enough for such prestigious outlets. Multiply this behavior across many thousands of scientists worldwide, and the stupidity of our collective action is nothing short of staggering.

Fast-forward 15 years and psychology remains hooked on JIF. When I applied for promotion at my own university in 2013, the promotion criteria—which were sensible in many other respects—stated that "[e]ach publication in the [applicant's] list should be accompanied by the impact factor of the

journal." Why? What possible information could such a discredited statistic possibly tell a promotion committee about my academic achievements or future prospects? Two years later when I examined a PhD at Maastricht University in the Netherlands I was confronted with an assessment form in which the student not only had to count the number of publications she had published (a questionable practice in itself), but was also required to add up the JIFs of those publications to produce a "summed impact factor."[12]

Evaluation of PhD thesis of: xxxxx xxxxx
by: [Prof. dr. C. Chambers]

PhD Length	*3.5 years*
OUTPUT STATISTICS	
Number of empirical chapters not published	2
Number of empirical chapters published	4
Number of first authored chapters	6
Summed impact factor of publications	20.124

Please rate output, skills, theoretical thinking and formal presentation (1 = below average; 2 = average; 3 = very good; 4 = exceptional; most scores in most theses will hover around 2)

OUTPUT (SCORE 1–4)	
Overall productivity/quantity of output	
Evaluation of summed impact factor	
Summed score	
SKILLS (SCORE 1–4)	
Experimental design	
Data acquisition methods	
Level of statistical analysis	
Computational approaches, modeling	
Summed score	
THEORETICAL THINKING (SCORE 1–4)	
Originality of ideas in all articles chapters	
Scholarly depth in all articles/chapters	
Scholarly depth of General Introduction	
Scholarly depth of General Discussion	
Summed score	

FORMAL PRESENTATION (SCORE 1–4)	
Quality of academic English writing	
Quality of figures *Summed score*	
TOTAL SCORE	
Total summed score	

Please judge which individual publications are of high impact (High impact corresponds to the top 5% of journals in the applicable research field. If applicable, write down in empty field the start page number[s] of the high impact publication[s].)

HIGH IMPACT PAPERS	
Number of publications of high impact	N/A

Psychology's ongoing obsession with JIF is revealed no more blatantly than in this perplexing PhD assessment from Maastricht University in the Netherlands—a university that has joined 13 other Dutch institutions in signing DORA. The form shown here, from 2015, requires the student to add up the JIFs for all articles they have published during their PhD to calculate a "summed impact factor." The examiner is then asked to evaluate the summed impact factor.

As an examiner I was then expected to offer an "evaluation of the summed impact factor." I found this requirement astonishing given that, in the previous year, all Dutch universities had signed the San Francisco Declaration on Research Assessment (DORA). DORA commits signatories to "not use journal-based metrics, such as Journal Impact Factors, as a surrogate measure of the quality of individual research articles, to assess an individual scientist's contributions, or in hiring, promotion, or funding decisions."[13] Soon after I submitted my assessment form, in which I pointed out the contradiction between signing DORA and using JIF to assess students, I received a reply from a senior university administrator explaining that the evaluation of PhD students according to their summed impact factors was a new mechanism being trialed.

> please note that this evaluation form has still a "pilot" status (it is not [university] broad, not faculty, but [departmental] policy), and the impact and publication rank issues will be evaluated /adjusted. . . . We also take care that critical issues detected in this pilot will not influence the evaluation of the thesis in any negative sense.

Despite the DORA serving as a national repudiation of JIF, it appears that individual departments within Dutch universities are free to create mechanisms that disregard it.

The JIF remains a popular metric despite no good evidence that it captures anything meaningful about the quality of science or scientists. If our attachment to JIF tells us anything, it is only how the sin of bean counting overwhelms our sense of reason. Brembs and colleagues conclude that "[m]uch like dowsing, homeopathy or astrology, journal rank [including judgments according to JIF] seems to appeal to subjective impressions of certain effects, but these effects disappear as soon as they are subjected to scientific scrutiny."[14] Structural biologist Stephen Curry, from Imperial College London, goes even further in condemning our collective foolishness:

> [A]stonishingly for a group that prizes its intelligence, [we] have acquired a dependency on a valuation system that is grounded in falsity. We spend our lives fretting about how high an impact factor we can attach to our published research because it has become such an important determinant in the award of the grants and promotions needed to advance a career. We submit to time-wasting and demoralising rounds of manuscript rejection, retarding the progress of science in the chase for a false measure of prestige.

Curry offers the following warning to anyone who uses JIF as an indicator of pretty much anything at all:

> —If you include journal impact factors in the list of publications in your *cv*, you are statistically illiterate.
> —If you are judging grant or promotion applications and find yourself scanning the applicant's publications, checking off the impact factors, you are statistically illiterate.
> —If you publish a journal that trumpets its impact factor in adverts or emails, you are statistically illiterate. (If you trumpet that impact factor to three decimal places, there is little hope for you.)
> —If you see someone else using impact factors and make no attempt at correction, you connive at statistical illiteracy.
> The stupid, it burns. Do you feel the heat?[15]

Unfortunately, it seems that most psychologists have yet to even get the message let alone feel any heat. Of the 840 organizations that have signed DORA in officially renouncing JIF, the Association for Psychological Science (US) is the only major professional group to have done so in psychology. Conspicuously absent are the American Psychological Association, the Psychonomic Society, and the British Psychological Society, as well as numerous other national and international psychological groups.

Even the Association for Psychological Science offers a strangely conflicted rejection of JIF. Four months after signing DORA, the then chief editor of its flagship journal, *Psychological Science*, Henry L Roediger III, wrote an insightful critique of JIF but concluded by saying: "I have been hard on [J]IF, but let me play devil's advocate and admit that [J]IF, even with its flaws, captures something real about journal quality." Contrary to DORA, Roediger even falls short of dismissing JIF as an indicator of individuals, leaving room for it at the table provided other factors are also considered: "Even if [J]IF values have some uses, when judging job candidates, tenure candidates, and grant proposals, the focus should be on the individual (or proposal), not just on [J]IF of journals where the candidate published."[16] To further muddy the waters, *Psychological Science* is promoted by its publisher as sporting "a citation ranking/impact factor that consistently places it in the top 10 psychology journals worldwide."[17] Even within one of the few psychological organizations to have signed DORA, it is clear that JIF retains a tight grip.

Wagging the Dog

Throughout my career I have been reasonably fortunate in attracting research funding, which means I sometimes find myself sitting in Kafkaesque meetings about university strategies to maximize so-called grant capture. Through the fog of management speak, a familiar scenario descends in which everyone in the room transforms into caricatures from the popular Irish sitcom *Father Ted*:

Professor A: "Our application success rates are too low!"
Professor B: "That's right, and we don't pull in as much grant funding as these other universities, here look at this list."

Professor C: "We need more grants!"
Professor A: "Yes, more grants! But how?"
Professor C: "We must *apply* for more grants! Then we will get more grants."
Professor B: "Brilliant! And let's also be sure to apply for more *expensive* grants. Then we'll rank higher on these arbitrary league tables that show which universities spend the most public money."
Professor A: "Oh yes, it's *very* important to spend more public money. How else will the world know how brilliant we are?"
[Everyone nods sagely]

Of course, the one thing that tends to be lacking from such strategy meetings is any talk of science, along with any recognition that grants are a means to an end rather than an end in themselves. And if I find myself rather wearily raising this issue, again, my comments are often met with an uncomprehending look or even an eye roll. *Who invited this guy?* Such is the nature of grant strategy—its purpose is not to ruminate on our scientific mission but to focus on how to hit funding targets, attract more money than our "rivals," hone the art of grantsmanship (the fabled skill of successful grant writing), and capitalize on the strategic priorities of funding agencies to ensure that we choose research questions that the agencies care most about and so will win us more money. Carl Bergstrom, a professor of biology at Washington University, sums up this mindset aptly: "I'm told that long ago, scientists pursued grants in order to do research, rather than vice versa."[18]

Why are research grants so important? At the level of departments and institutions, grant capture has become a banner of prestige and status—a way of showing the world that our particular organization is leading the way, ahead of the pack. At the level of individuals within institutions, the number of grants we receive (and their monetary value) feeds into decisions about who to hire and who to promote. At University College London, for instance, staff who apply for promotion are expected to show a "strong record of support through grants for research and travel," and in order to be promoted to the more senior academic ranks of reader or professor they are required to have "significant [i.e., lucrative and repeated] and sustained [i.e., long-term] success in obtaining research grants (and plans for further grant submissions)."[19] These requirements are not unusual among British and American universities. Despite the fact that, in most cases, grants are inputs

that pay for specific research projects, they have become regarded as outputs—fiscal quality indicators that are added to the assessment of research publications in determining the value of individual researchers. As psychologist Dorothy Bishop has observed, grants have become trophies on the shelf rather than investments to be balanced against returns:

> Many people seem to perceive a large grant as some kind of "prize," a perception reinforced by the tendency of the Times Higher Education and others to refer to "grant-winners"[20]. Yet funders do not give large grants as gestures of approval: the money is not some kind of windfall.[21]

But hang on, you might ask, is it entirely wrong to assess individual scientists or their institutions based on grant capture? Surely what makes an awarded grant an indicator of research quality is the fact that someone other than you felt your research was important enough to fund, right? While this is partially true, it is also misleading. What counts in science is the quality of the research that is actually conducted and published, not the hypothetical quality of research we *plan* to conduct, regardless of how impressive that plan may sound.[22]

Our system of academic promotion highlights the folly of rewarding scientists for plans rather than outcomes. Consider a scenario in which two psychologists are competing for one promotion opportunity. Jane applies with no major grants but a track record of high-quality publications. She can also show that her research has had a significant influence on fundamental theory, despite being unsupported by grant funding.[23] Now let's pit her against Mary, a researcher in the same field and at the same career level. Mary has a comparable publication record and research influence, but unlike Jane her research was supported by large public grants. Who do you think should be promoted? If you asked your next-door neighbor (or any economist for that matter) they would probably choose Jane and praise her efficiency and ingenuity—after all, conducting impactful projects without major funding requires a special degree of resourcefulness.

In our academic promotion system, however, Mary would wipe the floor with Jane.[24] Why? Because in judging a researcher's achievements we sum inputs (money spent) with outputs (research produced) rather than dividing them. And this, in turn, means that our system would value Mary over Jane for the sole reason that *Mary's research was more expensive*. Being awarded

a research grant is a fine achievement and something for which a researcher can feel justly proud. But grant capture should be completely irrelevant for assessing the contribution of researchers to science because a grant is not a contribution to science. Rewarding academics for acquiring grants is like deciding the winner of a football tournament based on which team spent more on boots.

Every so often the sin of bean counting reaches extraordinary heights. Since 2012, a number of major UK universities have set minimum funding targets in order for academics to merely retain their jobs. While none of these cases occurred specifically within psychology, they nevertheless hold important lessons for our field. In 2012, Queen Mary University of London was the first to announce that avoiding redundancy would require staff to acquire £50,000 per year of grant income as a lecturer, £65,000 as a senior lecturer, £80,000 as reader, and £100,000 as a professor.[25] At the Institute of Psychiatry at Kings College London, the criteria were even higher, revealed in a leaked internal document to reach £200,000 for professors.[26] At Queen Mary, these new rules were compounded by also requiring staff to publish a minimum number of so-called high-quality publications over a four-year period, with quality defined solely as publishing within outlets that have JIFs above 7.0. Such bare-faced bean counting prompted two Queen Mary academics, John Allen and Fanis Missirlis, to declare their university's policy "a triumph of vanity as sensible as selecting athletes on the basis of their brand of track suit."[27] Both researchers subsequently lost their jobs.

The Murky Mess of Academic Authorship

A successful career in psychology, as in others sciences, depends on more than simply chalking up a long list of grants and publications. Equally important are our individual contributions to published work and the resulting credit we claim, a point highlighted in section 6.1 of the position statement of the Committee on Publication Ethics, *Responsible Research Publication: International Standards for Authors*: "The research literature serves as a record not only of what has been discovered but also of who made the discovery."[28] As part of every piece of research there is a crucial negotiation between parties for attribution. Who was the main progenitor of the idea?

Who did the most work? Who analyzed the data? Who wrote (or will write) the manuscript? Did person X do enough to justify authorship as opposed to earning a thank-you in the acknowledgements? If X is included, do we have to also include Y? The list of questions goes on.

Careful negotiations about author attribution are vital because greater credit rightly belongs to researchers who made a greater scientific contribution. Beyond this point, however, everything about our system of attribution fails us miserably. The biggest problem is that rather than listing and quantifying the contributions made by researchers to an academic paper, we instead reduce this rich information to a crude heuristic: the published order of authorship. In psychology, order of authors works (mainly) like this—and because the rules are so murky and capricious the following points are highly caveated. The first-named author is usually the researcher who made the greatest intellectual contribution to the study and, again usually (but not always), the person who took responsibility for data analysis and much of the interpretation. Typically the first author is also expected to take the lead in writing the paper and coordinating the drafting process with the other coauthors. After the first author, the next most important position is the last author. The last author, or senior author, is usually the principal investigator—the top dog who made the study possible either by holding the grant that funded it or by supervising the student who conducted it.[29] If academic papers were movies, the first author would be the director while the last author would (usually) be both the producer and, ideally, the assistant director. In between the first and last positions lie the various grips and gaffers, and attribution here becomes even more nebulous. Second author is typically the next most coveted spot, held by someone who contributed significantly but slightly less than the first author. Between second and last lies the vast tundra of middle authors, whose actual contributions can vary enormously without anyone—even the authors—really knowing why they are in a particular position. For example, the fourth author of a seven-author paper might be a specialist statistician who designed and implemented a vital part of the analysis, or it might be an intern who collected some data on a quiet weekend, or even a collaborator who shared a bit of code early on and never saw the paper again. Most psychology journals don't list the specific contributions that each author made to the design, implementation, analysis, interpretation, and writing

of published papers; they simply list the order and leave it for the community to guess.

The second major problem is that factors other than scientific contribution routinely influence the degree of attribution. In some areas of neuropsychology and psychiatry, it is standard practice for the clinician who supplied patient volunteers to be offered coauthorship as compensation, despite this person making no contribution to any other aspect of the research. Similar customs apply for principal investigators (lab leaders), who claim authorship on papers emerging from their lab even when they had little intellectual or practical involvement in the work. Even more egregious are various forms of "gift authorship." Sometimes this happens by falsely inflating the contribution of an existing team member on the grounds that, in order to provide an immediate career boost, one coauthor needs a higher level of credit than another. Other times, authorship is offered to colleagues either as a social gesture (to maintain a working friendship or in anticipation of quid pro quo) or in the case of heavyweight colleagues as a form of celebrity endorsement to bias the journal editor into accepting the paper. All such practices violate established guidelines for authorship outlined by the American Psychological Association (APA) and the International Committee of Medical Journal Editors (ICMJE). The APA specifically notes that authorship is not necessarily warranted for "providing funding or resources, mentorship, or contributing research but not helping with the publication itself."[30] The ICMJE further outlines four necessary and sufficient conditions that must be met by any researcher wishing to claim coauthorship:

- Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND
- Drafting the work or revising it critically for important intellectual content; AND
- Final approval of the version to be published; AND
- Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.[31]

A third problem is that attribution by rank order makes it impossible for authors to make an equal contribution, despite this being a frequent

possibility in any group-led scientific enterprise. Under a ranking system, the only way to represent equal attribution is through a clumsy asterisk next to two (or more) names in the byline, with a footnote stating that "these authors contributed equally to the work." Such footnotes are easily missed when reading a paper and are not stated at all when a paper is cited. Rank order requires that one author is always listed before another and, for better or worse, that author will be seen as the more important contributor.

The ethics of academic authorship are impossible to police and are routinely flouted because the culture of bean counting in psychology rewards academics who publish a high volume of first- and last-authored publications. If you are a junior researcher in psychology you will want to tally up a high quantity of first-authorships, ideally in journals that are prestigious or with high JIFs. As you progress into research leadership and emerge as a "lab head," it becomes more important to cultivate a rich crop of last authorships as an indicator of seniority. In this way, we can see how the sin of bean counting manifests not simply as a ledger of articles published, but as the quantity of articles published with specific attributions, over and above the scientific quality and importance of the work, and ignoring the murky imprecision of rank authorship.

I myself am guilty of fueling this obsession with authorial bean counting. In 2012, I wrote a blogpost that offered 25 tips for early career researchers. Regrettably, some of my advice emphasized the quantity of publications with particular attributions rather than the quality of science being performed:

> Aim to publish every year. . . . The rule of thumb in psychology and cognitive neuroscience is to publish four good papers per year, but this can vary. Whatever happens, be sure to publish *something* every year. If your experiments are slow or not yet producing publishable data, then publish a review paper. . . . Aim for as many first-authorships as possible.[32]

In response, Dorothy Bishop quite rightly called me out for bean counting, writing in a comment below my article:

> I am uneasy with the recommendation to publish 4 papers per year, and to write up as much as possible of your thesis. I think you should

> publish when you have done something important enough to communicate to the rest of the world. . . . I'd rather see a c.v. with one really carefully-done, thoughtful, original paper every year, than one with ten papers per annum, none of which is memorable in any way. It's a balance that's not always easy to get right, but my advice is, remember that if you write stuff, editors and reviewers have to read it, and they may form a negative impression of you if it's trivial or pointless stuff, or if you have clearly been chopping up a decent study into minimal publishable units.[33]

This gave me pause for thought. Why *had* I offered this advice? In my reply, I diverted the blame to decision makers, as many researchers who offer such suggestions do. It wasn't my fault, I implored, because I was told so by the powers that be. Don't shoot the messenger, and so forth. It was a weak defense:

> This was the advice I received time and again from the BBSRC fellowships committee; in fact I remember a slide from the induction session of my David Phillips fellowship in 2006 which stated "Publications: ensure a balance between *Nature* and potboilers." And when I went for my mid-term assessment, they made a point of warning me that because I was transferring institutions, I should guard against a drop in the *quantity* of publications. . . . from my experience, at least, quantity of output is still regarded as an important gauge of individual success in science. I'm not saying it should be.[34]

It is a curious experience to reread one's own words and see them as fundamentally flawed. What I realize now is that in any resource-limited system, quantity and quality lie at opposite ends of a seesaw. This means that by weighing the value of scientists *in any way* according to the quantity of journal articles they produce, we necessarily devalue quality. At its extreme, valuing quantity encourages salami slicing of research into what are known as "minimal publishable units"—the smallest amount of research that can be accepted in a journal and thus count in a CV—which in turn floods the peer-reviewed literature with an overabundance of hurried, small, and oversold studies. Instead we should be rewarding scientists for publishing a fewer number of higher quality studies that are more theoretically important, larger, and more methodologically rigorous. Developmental psy-

chologist Uta Frith has referred to this revisionist philosophy as "slow science," drawing an analogy to cooking.

> Slow Food and Slow Science is not slowing down for its own sake, but increasing quality. Slow science means getting into the nitty gritty, just as the podding of fresh peas with your fingers is part of the production of a high quality meal. Science is a slow, steady, methodical process, and scientists should not be expected to provide quick fixes for society's problems.[35]

Frith's argument is compelling, but what about the counterargument that a complete deemphasis on the quantity of publications would invite exploitation by researchers, and even laziness? How could two researchers be compared if their work was judged of similar quality but one had simply produced a greater amount than the other? Should they be judged equivalently? The answer is that assessments of quality can take into account quantity, but only indirectly: producing a larger volume of research would increase the overall quality of a researcher's contribution to science only if the contribution to theory or practice rises in proportion with the amount of work. On the other hand, it is possible that for some researchers, quality may increase only slightly, or not at all, with quantity—for instance, a collection of multiple papers reporting low-powered experiments could be judged as the same or lower quality than a single paper describing a rigorous, high-powered investigation. To avoid perverse incentives, the key requirement in any quality-based system is that the assessment of individuals, or comparisons between them, must be made independently of the sheer number of publications.

Roads to Somewhere

In this chapter we have considered three ill-judged ways of assessing the quality of psychological science and scientists—journal impact factors, grant awards, and authorship rank. So are metrics and quality indicators completely useless? Should we abandon them entirely?

No. It is natural to be drawn toward citation metrics. We all want our work to be influential, and we take pleasure in it being cited because it means someone read it and, with luck, found it useful.[36] But there is a criti-

cal difference between things that make us feel good and the criteria we should use to assess and value other scientists. Recognizing that difference and applying it requires mental discipline. Metrics like JIF are meaningless as indicators of scientific quality, but citations themselves at the level of individual articles—while not indicators of quality—are suggestive of influence and interest-value.[37] There is no shame in paying attention to these features of our academic landscape provided we see them for what they are and, crucially, for what they are not. The lazy worship of metrics should never replace informed expert judgment. Only expert appraisal can render authoritative judgments on scientific quality (and even then it can often be wrong). We allow metrics to supplant our subjective powers at our own peril.

The same principle applies to grant funding. It is perfectly understandable for scientists to celebrate the award of a grant that enables them to pursue research that would otherwise have remained out of reach. Being awarded funding carries its own intrinsic rewards—as researchers we now get the resources to address an exciting and important question, and the money in the grant can allow us to buy our time out of teaching and administrative responsibilities within our university departments. As with citation metrics, however, grant income should never be used to assess the quality or potential of researchers because a grant is a mission plan, not a completed mission. If grants are considered at all, they should be weighed *against* research outputs: assessors should ask whether the scientific contribution of this work was in proportion with the resources the researcher consumed. Assessors should not simply reward the act of consumption.

Finally, what do we do about the convoluted mess that is academic authorship? In the final chapter we will consider an alternative *contributorship model* that borrows certain features from the physical sciences. Under a contributorship model, the order of authorship is irrelevant—all that matters is the work that authors actually undertook to earn their positions on a paper. Authors thus earn credit for the extent of their contribution to different roles, whether originator of the theory or hypothesis, developer of methods, gatherer of data, and so on.

The sin of bean counting exists because we lack a robust way of assessing the quality of scientific output in the short to the medium term. How do we weigh up the scientific contributions of different individuals in the competition for jobs and funding? How do we decide if a completed grant was

successful? Do we check whether independent scientists replicated the funded studies, or at least that the studies were conducted rigorously and are therefore replicable? Until psychology develops a proper mechanism for attributing credit and assessing genuine research quality then bean counting will continue to reign supreme.

CHAPTER 8

Redemption

The method of science, as stodgy and grumpy as it may seem, is far more important than the findings of science.
—Carl Sagan, 1995

Throughout the last seven chapters we have seen how psychological science falls prey to a number of cultural sins, including bias, hidden analytic flexibility, unreliability, lack of data sharing, vulnerability to fraud, failure to make published work publicly available, and an obsession with bean counting. Left unchecked, these sins not only hold back discovery and practical applications of psychology; they also pose an existential threat to the discipline itself. In an increasingly competitive world—a world in which public spending on science is under intense pressure—psychology can ill afford to bury its head in the sand and hope these failings will be ignored.

We can anticipate two objections to this conclusion. On the one hand, defenders of the status quo will argue that none of these failings are unique to psychology. Questionable research practices and indifference toward replication are present across much of the life and social sciences, and publication bias is ubiquitous to varying degrees across all science. Lack of data sharing, barrier-based publishing, fraud, and the worship of impact factors threaten every field, not just psychology, and not even just science but the humanities and social sciences too. So, why should we focus on psychology? And why should it fall on psychologists to "fix" science?

The answer is simply this: as psychologists we must take responsibility for the quality of the research we produce, regardless of what may be happening in other fields. I focus on psychology because psychology is where I live. If the roof of my house is leaking and threatening to collapse, I don't look at my neighbor's house and say "They have the same problem so I don't need to worry about mine." I fix my roof, and, in doing so, maybe I can develop a solution that helps fix other roofs as well. But if everyone passes the buck then repair never happens. The diffusion of responsibility is a key reason why reform in psychology has been so slow, despite many of the flaws identified here being known for decades.

At the opposite end of the spectrum lie those who will take these sins as proof that psychology isn't a science and should therefore be abandoned. This is equally false. Precisely what makes an area of study a science is a vexed question that has provoked strong and varied reactions from philosophers for over a century. To adopt a Popperian perspective, a scientific discipline can be said to be one in which the phenomenon under investigation is quantifiable, the hypothesis testable, the experiment repeatable, and the

theory falsifiable. Much psychological research clearly meets all these conditions even when its practitioners—psychological scientists—fail to meet the highest standards in every area.[1] For example, most studies in quantitative psychological fields (such as psychophysics, cognitive psychology, experimental psychology, and social psychology) include objective measurements of behavior that are repeatable in principle. It's just that few experiments are exactly repeated and many findings fail to replicate when they are, which can be explained by practices that are weak but nevertheless scientific. In the same way, most hypotheses are precise enough to be falsifiable even if researchers succumb to confirmation bias or hindsight bias, but again these failings fall within the spectrum of scientific practices. The susceptibility of psychology to these problems is comparable with other areas of biomedical research, which have been shown to suffer from low reproducibility, research bias, and publication bias.[2] And lest the reader is still tempted to cast psychology into the fire, we should remind ourselves that in spite of the severity of these problems we have nevertheless assembled a remarkable vault of knowledge about the mind and brain. Over the last century, hundreds of robust psychological phenomena have been established and fed into the development of theory. These insights have changed the way we think about human cognition, behavior and society.

Advances aside, there is no denying that psychology is less robust than many other sciences, in particular compared with the physical sciences. Fields such as physics and chemistry are less prone to bias, place more emphasis on replication, offer more precise and falsifiable theories, and manifest more open practices. Psychology instead shares characteristics with Richard Feynman's famous definition of "cargo cult science." Returning to his 1974 essay, which we discussed in chapter 3:

> In the South Seas there is a cargo cult of people. During the war they saw airplanes land with lots of good materials, and they want the same thing to happen now. So they've arranged to imitate things like runways, to put fires along the sides of the runways, to make a wooden hut for a man to sit in, with two wooden pieces on his head like headphones and bars of bamboo sticking out like antennas—he's the controller—and they wait for the airplanes to land. They're doing everything right. The form is perfect. It looks exactly the way it looked before. But it doesn't work. No airplanes land. So I call these things cargo cult science, because they follow all the apparent pre-

cepts and forms of scientific investigation, but they're missing something essential, because the planes don't land.

The good news for psychology is that the planes do land. The bad news is that they don't come as often as they should because we have failed to heed Feynman's warning that "you must not fool yourself, and you are the easiest person to fool." In this final chapter we will discover how psychology can meet Feynman's challenge and, in doing so, fortify and enrich its mission.

Solving the Sins of Bias and Hidden Flexibility

In chapters 1 and 2 we saw how biased research practices weaken the published evidence base in psychology. Confirmation bias sabotages various stages of the scientific process, from publication bias in favor of positive and novel results to the reliance on "conceptual replication" that can confirm but never refute previous findings. We also saw how hindsight bias manifests through the questionable practice of Hypothesizing After Results are Known (HARKing) in which the researcher either knowingly or unknowingly presents an unexpected result as though it was predicted all along. Meanwhile, hidden flexibility in research analyses, including *p*-hacking, inflates the rate of false discoveries. We have already discussed some of the "nudge" incentives that can help reduce bias. Badges for rewarding open practices, methodological checklists, and disclosure statements are all valuable and useful initiatives. However there are fundamental limitations in what they can achieve because, while they incentivize transparency and unbiased practices, they don't actually prevent them. We may never be able to eliminate bias altogether from human nature, but there is one sure way to immunize ourselves against its consequences, and that is peer-reviewed study preregistration.

Registered Reports: A Vaccine against Bias

My earliest experience of publication bias came in 1999 with the rejection of my first scientific paper.[3] "The results are only moderately interesting," chided an anonymous reviewer. "The methods are solid but the findings

cepts and forms of scientific investigation, but they're missing something essential, because the planes don't land.

The good news for psychology is that the planes do land. The bad news is that they don't come as often as they should because we have failed to heed Feynman's warning that "you must not fool yourself, and you are the easiest person to fool." In this final chapter we will discover how psychology can meet Feynman's challenge and, in doing so, fortify and enrich its mission.

Solving the Sins of Bias and Hidden Flexibility

In chapters 1 and 2 we saw how biased research practices weaken the published evidence base in psychology. Confirmation bias sabotages various stages of the scientific process, from publication bias in favor of positive and novel results to the reliance on "conceptual replication" that can confirm but never refute previous findings. We also saw how hindsight bias manifests through the questionable practice of Hypothesizing After Results are Known (HARKing) in which the researcher either knowingly or unknowingly presents an unexpected result as though it was predicted all along. Meanwhile, hidden flexibility in research analyses, including *p*-hacking, inflates the rate of false discoveries. We have already discussed some of the "nudge" incentives that can help reduce bias. Badges for rewarding open practices, methodological checklists, and disclosure statements are all valuable and useful initiatives. However there are fundamental limitations in what they can achieve because, while they incentivize transparency and unbiased practices, they don't actually prevent them. We may never be able to eliminate bias altogether from human nature, but there is one sure way to immunize ourselves against its consequences, and that is peer-reviewed study preregistration.

Registered Reports: A Vaccine against Bias

My earliest experience of publication bias came in 1999 with the rejection of my first scientific paper.[3] "The results are only moderately interesting," chided an anonymous reviewer. "The methods are solid but the findings

are not very important," said another. "We can only publish the most novel studies," declared the editor as he frog-marched me and my boring paper to the door. The experience was not what I expected. I was 22, a fresh-faced PhD student in his first year, instilled with the belief that science was a process of objective truth seeking. The very last thing I expected was that a scientific journal—and one so highly respected in my field—would reject a piece of theoretically and methodologically sound research based on the *results*. It just didn't make sense and seemed incredibly unfair. How could we be expected to control the results? Didn't this just turn publishing into a lottery? Surely there had to be some mistake?

My supervisor wasn't surprised. Although he had approved a draft of the manuscript before submission, he had harbored reservations about the way it was presented. "It didn't really tell a story," he said. "I could see what you were trying to do, but you have to give the reviewers a narrative—something they can get excited about—or they will find your paper boring and reject it." The conversation was formative in my training. "But the results are the results," I exclaimed. "Shouldn't we just let the data tell the story?" He shook his head. "That's not how science works, Chris. Data don't tell stories, scientists tell stories."

Data don't tell stories, scientists tell stories.

Over the next decade I heeded those words. We told all kinds of stories based on data. Big stories led to papers in zeitgeist journals like *Nature Neuroscience*, *Neuron*, and *PNAS*. Smaller stories led us to solid specialist journals. We'd carefully design and conduct experiments and then weave compelling narratives out of the results. It was fun and we published well. We were awarded grants. I got a career. Everybody won.

Or did they? Every paper had a story, but was that story the most unbiased representation of the data? Of course not. We would gloss over imperfections or noisy results that were hard to explain, burying the nuances in "supplementary information" or not mentioning them at all. I'm certain that I committed the sin of hidden flexibility many times, and no doubt at an unconscious level even more than I realize. Every paper had a pithy logline and a tale to tell. As hard as we worked on the science—and we worked very hard to design experiments that were theoretically meaningful—we worked tirelessly on the storytelling.

The strategy worked. In 2006 I was awarded a fellowship that allowed me to move to the UK and establish my own research team at a major Lon-

don university. Initially I practiced the same tried-and-true formula, one that my new institute already embraced: novel experiment + great results + great storytelling = publication in a prestigious journal. To succeed you needed every element in place; if even one was missing or below par then your study would end up in a lower-ranking specialist journal or the file drawer. By now, however, this style of science was starting to grate on me. I had always found the publish-or-perish culture unappealing, and the pressure at my new institute was higher than ever. As scientists, the one part of an experiment we were supposed (in theory) to relinquish control over was the results. To teach this but to nevertheless pin success on "good results" was a devil's temptation toward bias, questionable research practices, and fraud. I was tired of watching colleagues analyzing their data a hundred different ways, praying like gamblers at a roulette wheel for signs of statistical significance. I was fed up with the inexorable analysis and reanalysis of brain imaging data until a publishable story emerged from an underpowered design. I also became dubious about the robustness of my own work. As a friend dryly observed, "You guys do high-impact work, but you're like magpies. You do one study and move on to something else without ever following it up." He was right, of course. "That's the game," I admitted. "Why would anyone waste time doing the same experiment twice when no funder will pay for it and no top journal will publish it? And why would you take the risk of failing to replicate yourself?" Talk about shooting yourself in the foot.

As my doubts grew, the idealism of my younger self began to reassert itself. My institute was packed to the rafters with brilliant people, but it felt like the scientific equivalent of an elite telemarketing company. Walking in the front door each morning you would face off against a wall-mounted screen listing this week's "top sellers"—the roll call of who published what in which prestigious journal. *Nature Neuroscience*. *Current Biology*. *Science*. The last-author credit would always belong to one of the professors, with the first-author slot filled by a tireless protégé who barely left the building. If your name wasn't on that list (and mine usually wasn't) you felt small, inadequate, an imposter. You pushed yourself harder. You pushed your staff and students harder. As the director of the institute at the time—and a valued mentor—once told me: "This place is powered by appetite. You keep the young researchers hungry. You keep them on fixed-term contracts with uncertain futures, and you place them in competition with the world and

each other. The result is high-octane science." I was never convinced that he really believed in this philosophy, but it was the reality within those walls.

After two years in London I left for a more stable academic career, and admittedly a more relaxed professional lifestyle. For several years I continued cranking the handle before a series of events reawakened my idealism with a jolt. In 2011 we had a paper rejected by the *Journal of Cognitive Neuroscience* for the main reason that because one of the experiments was a close replication of the other, the study and results weren't considered sufficiently novel or important to be publishable. One of the reviewers even told the editor in a private comment (obtained by us only after we had unsuccessfully appealed the rejection): "The methods are strong, but the outcome produced results that aren't particularly groundbreaking."

A year later I snapped. In September 2012 we received a rejection letter from the specialist journal *Neuropsychologia*, in which our paper was declined because one of the analyses (and not even a test that was important for the conclusions) returned a statistically nonsignificant outcome. We hadn't glossed over it; we hadn't massaged it away by *p*-hacking or HARKing; we had simply reported it honestly and transparently. It was like 1999 all over again. Unpublishable results. Boring study. Reject. A few days later I sat down and wrote a letter to the editors of *Neuropsychologia*, which I also posted on my personal blog.[4] I thanked them for their work on our paper and made it clear that what I was about to say wasn't personal. I had reviewed for the journal and published in it many times, but I was now severing all ties. "My real problem is that by rejecting papers based on imperfect results, *Neuropsychologia* reinforces bad scientific practice and promotes false discoveries. . . . For this reason, my main purpose in writing is to inform you that I will no longer be submitting manuscripts to *Neuropsychologia* or reviewing them. I will be encouraging my colleagues similarly."

In the meantime I began to form an idea for a solution. I had been following the work of blogger Neuroskeptic, who for several years had been proposing study preregistration as a way to improve transparency and reproducibility.[5] In his articles, and in the comments below them, you could always find a vigorous debate about the potential benefits and drawbacks of preregistration. Requiring authors to prespecify their hypotheses and analysis plans held great promise for reducing a wide range of questionable research practices, such as *p*-hacking and HARKing. Furthermore, if jour-

nals could somehow be convinced to accept preregistered studies without knowledge of the results then it would also prevent publication bias and eliminate the incentive for authors to engage in questionable practices in the first place. It seemed to me a brilliant solution to many problems. Suddenly there would be no further need for excessive storytelling, no need to gloss over inconsistencies and "messiness" of data. But how could you persuade authors and journals to actually do this? Should it be mandatory? Would the process be too cumbersome and bureaucratic? There were dozens of unanswered questions, but I found the debates fascinating, and I had an idea. I just needed a journal to try it in.

An opportunity soon came. A few weeks after publishing my open letter to *Neuropsychologia*, the chief editor of *Cortex*, Sergio Della Sala, invited me to join his editorial board. I accepted and immediately set to work on a proposal for a new type of empirical article called a Registered Report in which study protocols (including introduction and methods) would be reviewed *before* authors gathered their data.[6] Registered Reports stem from the simple philosophy that the publishable quality of science should be judged according to the importance of the research question and rigor of the methodology, and never based on whether or not the hypothesis was supported. This wasn't a new idea and certainly wasn't mine. In addition to Neuroskeptic's blog posts calling for preregistration, there had been several proposals within the last 50 years for such a publishing mechanism. As far back as 1966, psychologist Robert Rosenthal wrote:

> What we may need is a system for evaluating research based only on the procedures employed. If the procedures are judged appropriate, sensible, and sufficiently rigorous to permit conclusions from the results, the research cannot then be judged inconclusive on the basis of the results and rejected by the referees or editors. Whether the procedures were adequate would be judged independently of the outcome.[7]

A few years later, in 1970, G. William Walster and T. Anne Cleary from the University of Wisconsin offered the same idea:

> In proposing this alternative policy, we argue that all decisions involving the treatment of data should be considered design decisions. Then, since the decision to publish the results of a study is particular

> treatment of data, it follows that the same limitations should be imposed on publication decisions as are imposed on all designs. When one views publication in this way, it becomes immediately clear that a specific change should be made in current policy. There is a cardinal rule in experimental design that any decision regarding the treatment of data must be made prior to an inspection of the data. If this rule is extended to publication decisions, it follows that when an article is submitted to a journal for review, the results should be withheld. This would insure that the decision to publish, or not to publish, would be unrelated to the outcome of the research.[8]

Neither Rosenthal's nor Walster and Cleary's proposals were ever implemented, but perhaps Registered Reports could be our chance to finally do so. In the model I had in mind, protocols considered scientifically important and robust and that met strict guidelines for prospective rigor would be offered "in principle acceptance." The journal would then commit to publishing the outcomes regardless of how the results turned out, provided the authors adhered to their preregistered protocol, that various prespecified quality checks were passed, and that the conclusions were based on the evidence obtained.

Three days later my proposal was ready for feedback. I distributed it to the editorial board, and at the same time I also posted it as an open letter on my blog.[9] I knew that publishing an open letter would be controversial as journals are accustomed to considering such proposals behind closed doors rather than in sight of the public. But I had resolved to take an open route for two reasons. My foremost concern was that the proposal might have a glaring flaw that I hadn't considered, and the best way to find that out was to crowd source critical feedback from the wider community, including those who had commented on Neuroskeptic's earlier proposals. And secondly, I wanted the journal—and me—to be held accountable for whatever decision it reached about Registered Reports. I had heard on the grapevine that similar ideas had been mooted in the past at other journals and binned by conservative boards amid concerns that accepting papers before results could force the journals to publish negative or inconclusive outcomes. And even though I had known Sergio Della Sala for many years and respected him, the wider *Cortex* editorial board was an unknown quantity. I had no idea how the board would react in private, but going public pro-

vided a crucial test of the idea. If the journal decided to reject Registered Reports then its reasons would need to be sufficiently defensible to survive public scrutiny. And if the journal adopted them then the community, having been involved since the beginning, could play a role in shaping the initiative. In one case we would learn something important; in the other we would hopefully *do* something important.[10]

Within a few days my open letter had attracted thousands of views and dozens of comments below the line. The feedback contained many constructive suggestions for improvement, most of which I integrated into the proposal. However my strategy divided the *Cortex* editorial board. Some editors were supportive of Registered Reports, but many were not, even though none of the strongest opponents ever expressed their views to me personally. Several board members also felt that I had risen above my station in proposing the initiative in such a public way, just a few days after being invited to join the editorial board.

Their response was understandable. I knew that the open letter would seem like a coup to some, but I decided that the ire of a few editors was a small price to pay for exposing Registered Reports to the crucible of public opinion and giving it the best possible chance of a fair hearing. Fortunately, the chief editor was strongly in favor. A month later, in November 2012, *Cortex* became the first journal to approve Registered Reports. We assembled an editorial subcommittee to handle submissions and prepared for the launch in May 2013.[11]

How do Registered Reports work? Unlike conventional papers where peer review happens after the entire study is finished and written up, here the review process is split into two stages (see figure 8.1). At Stage 1, authors submit an introduction, proposed methods and the analysis plan *before* they have collected data. These are initially triaged for scientific significance, clarity, and adherence to specific Stage 1 review criteria. I took a lot of time to shape these criteria based on ideas and feedback from Neuroskeptic and many others. Editors and reviewers at Stage 1 assess:

1. The significance of the research question(s)
2. The logic, rationale, and plausibility of the proposed hypotheses
3. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis)
4. Whether the clarity and degree of methodological detail would

be sufficient to replicate exactly the proposed experimental procedures and analysis pipeline

5. Whether the authors provide a sufficiently clear and detailed description of the methods to prevent undisclosed flexibility in the experimental procedures or analysis pipeline
6. Whether the authors have considered sufficient outcome-neutral conditions (e.g., absence of floor or ceiling effects; positive controls) for ensuring that the results obtained are able to test the stated hypotheses

Let's take a moment to explore these points in more detail. The first two criteria are designed to test the scientific credibility of the proposal. Is the research question important? Do the hypotheses arise logically from the literature? The third criterion tests whether the proposed methods are rigorous and realistic, with particular emphasis on statistical power. As we saw in chapter 3, underpowered experiments are common in psychology and neuroscience, increasing the rate of false negatives and false positives. At *Cortex* we decided to set a minimum statistical power of 0.9, meaning that the experiment must have a sufficiently large sample that all statistical hypothesis tests have a 90 percent chance of correctly rejecting a false null hypothesis. We also invite authors to consider alternative Bayesian sampling and hypothesis-testing methods in which statistical power is irrelevant and data are simply acquired until a sufficiently decisive answer is reached.

The fourth and fifth criteria focus on the extent of detail in the proposed methods. Are the study procedures elaborate enough to provide a replication "recipe" that could be repeated by other researchers? This takes into account the major cause of unreliability in psychology (as discussed in chapter 3), that method sections are often too vague to permit replication. In addition, are the proposed analyses sufficiently precise to prevent "wriggle room" that could allow authors to either consciously or unconsciously exploit researcher degrees of freedom such as *p*-hacking? Finally, the sixth criterion requires authors to specify, in advance, what quality checks and positive controls are required in order for the study to provide a fair test of their hypothesis.[12] An experiment might produce meaningless results because equipment was wrongly calibrated, or because behavioral performance from the participants was at ceiling or floor levels, or because a

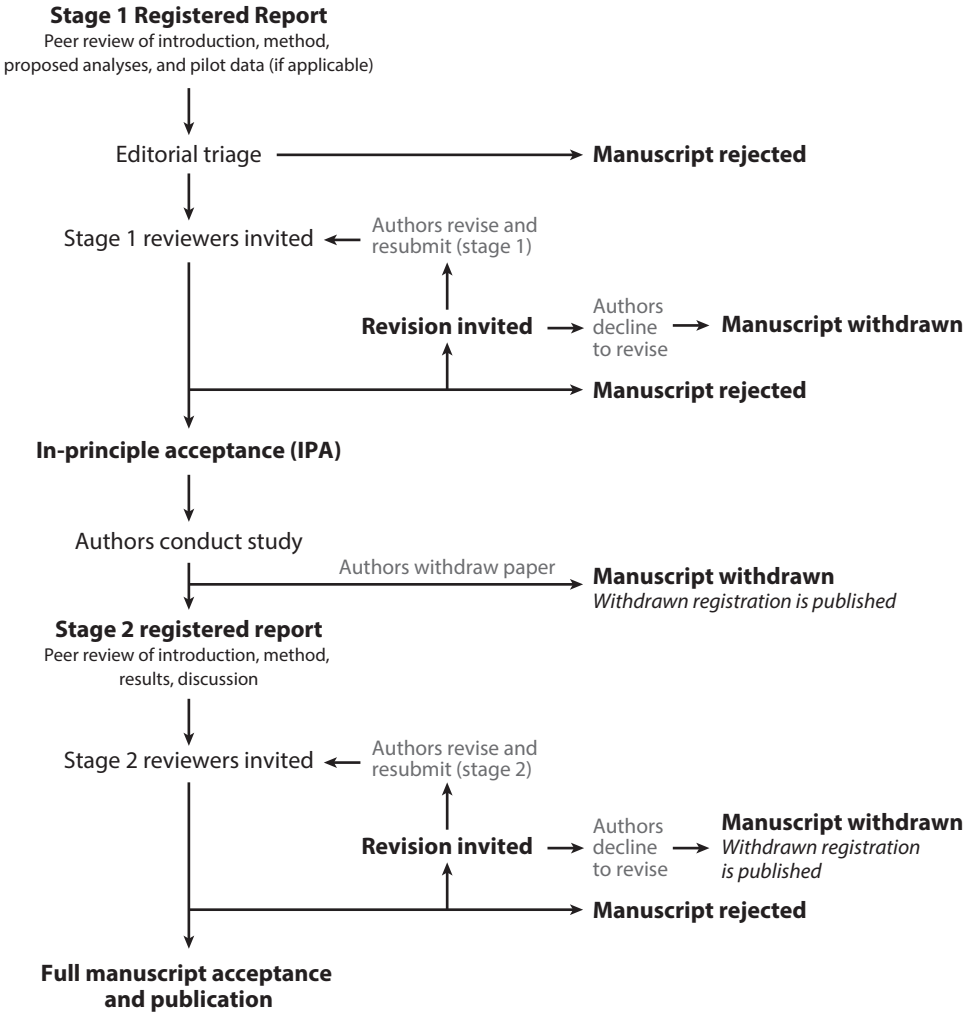


FIGURE 8.1. The submission pipeline and workflow for Registered Reports at *Cortex* and several other journals.

condition with a known effect failed (i.e., a reality check or "positive control"). To avoid publication bias, any such tests would have to be outcome-neutral, which is to say they must be independent of the study hypotheses. Keeping these tests independent of hypotheses prevents the common bias, seen in status quo publishing, of authors and reviewers using positive or negative outcomes to decide whether or not an experiment "worked," and to reject or accept it accordingly.

If the submitted manuscript passes editorial triage, it is then it is sent for in-depth peer review where experts assess the proposed study rationale and methods, testing it against the Stage 1 criteria. Following a favorable assessment, which might require revision of the protocol, the journal can then offer the paper "in-principle acceptance" or IPA. Only once IPA is obtained can researchers then implement the preregistered experiment. Following data collection and analysis, they resubmit a Stage 2 manuscript that includes the introduction and methods from the original submission plus the results and discussion. The results section of the completed manuscript must include the outcome of all preregistered analyses. Crucially, however, it can also include any additional unregistered analyses provided they are clearly distinguished and identified as exploratory (post hoc) rather than confirmatory (preregistered). The authors are also required to deposit their data in a publicly accessible archive, addressing the sin of data hoarding. At Stage 2 the reviewers, who are ideally the same as at Stage 1, consider the following criteria:

1. Whether the data are able to test the authors' proposed hypotheses by passing the approved outcome-neutral criteria (such as absence of floor and ceiling effects)
2. Whether the introduction, rationale, and stated hypotheses are the same as the approved Stage 1 submission (required)
3. Whether the authors adhered precisely to the registered experimental procedures
4. Whether any unregistered post hoc analyses added by the authors are justified, methodologically sound, and informative
5. Whether the authors' conclusions are justified given the data

The first of these criteria assesses whether the experiment passed the preapproved quality checks and positive controls. For example, if the study was on the effect of alcohol on cognitive function, did the authors confirm that the alcohol had the necessary intoxicating effect through an established questionnaire or other measure? If this positive control were to fail, then this would suggest that the intervention was administered incorrectly, and thus that the hypothesis was not properly tested. The second and third criteria check for consistency between the introduction and methods sections of the preregistered protocol with the same sections in the Stage 2 manuscript. Finally, the fourth and fifth criteria check that any additional explor-

atory analyses, and the overall study conclusions, are sensible. The finished article is published only after this process is complete.[13]

Note what is missing from the Stage 2 review criteria. There is no assessment of impact, novelty, or originality of results. There is no consideration of how conclusive the results happen to be. There is no weight placed on whether or not the experimental hypothesis was supported. Such considerations may be relevant for scientists in deciding whether a study is exciting or newsworthy, but they tell us nothing about scientific quality or the longer-term contribution of the study. For Registered Reports, the outcomes of hypothesis tests are irrelevant to whether a piece of scientific research meets a sufficiently high standard to warrant publication.

At around the time that *Cortex* launched Registered Reports in May 2013, similar initiatives began popping up at *Perspectives on Psychological Science* and *Attention, Perception, and Psychophysics*.[14] My impression was that preregistration was fast transforming from theory into reality, but I also became aware that some journals were quietly shelving the idea without transparent debate. Those of us working to promote Registered Reports were hearing that chief editors, petitioned behind closed doors to adopt the initiative, were rejecting it on the grounds that it would lead to the publication of negative findings that might be cited less than standard articles, leading to a consequent drop in the journal's impact factor. I found this worship of impact factors deeply disappointing, and it reinforced my belief that closed-door politics are the antithesis of rational decision making (and are probably the reason that Registered Reports had gone nowhere since the 1960s).

My colleague Marcus Munafò and I decided that something needed to be done or the initiative could be killed by the sin of bean counting before it got started. Three days after the *Cortex* launch we met at a pub in Bristol and devised a plan. Over the next month we would assemble dozens of scientists and members of journal editorial boards, and, in June 2013, we published a joint open letter article in the *Guardian* headlined "Trust in Science Would Be Improved by Study Pre-registration."[15] The article, which was eventually signed by more than 80 senior academics, called for all life science journals to offer Registered Reports as a new option for authors. Registered Reports were not going to go quietly into the night.

The storm that followed was astonishing. The publicity of the *Guardian* opened the debate to a much wider community of researchers, and two opposing camps took shape. On one side were the reformers pushing for

greater transparency and reproducibility; they were generally in favor of Registered Reports being adopted, if only to find out how the articles it produced compared with standard publishing. On the opposite side was a rearguard of (often powerful) traditionalists who argued that preregistration would "put science in chains."[16] I found these criticisms puzzling. Many were based on what appeared to be simple misunderstandings or elementary errors about the initiative. Had these people even read what we had written? Other reactions struck me as disingenuous misrepresentations that appeared to have no purpose except to derail the initiative and preserve the status quo. What follows are some of the main objections that emerged, my personal interpretation of the objection, and a longer explanation in each case:

- *Registered Reports prevent exploration of data and curb scientific creativity.* **Verdict: False.** This is perhaps the greatest misconception of the initiative. Authors of Registered Reports are welcome to perform unregistered exploratory analyses with as much creativity as they can muster, just as they would with standard publishing. The only requirement is that such analyses are labeled transparently as exploratory rather than being presented as confirmatory. I was particularly surprised at how many traditionalists clung (and still cling) to this argument, despite exploratory analyses being formally invited as part of Registered Reports policies at all adopting journals. It appears that some traditionalists not only want the freedom to conduct exploratory analyses (which Registered Reports explicitly welcome), but also want to be able to present those exploratory analyses as confirmatory hypothesis testing. I reached the rather unsettling conclusion that the traditionalists who continued to oppose Registered Reports on these grounds simply wanted the freedom to HARK, and because HARKing is socially unacceptable the opponents had no choice but to argue that Registered Reports would somehow prevent exploration.
- *Registered Reports could lead to the denigration of exploratory, observational research.* **Verdict: Unsubstantiated and probably false.** This is a subtler version of the first counterargument. The concern here is that developing a more robust mechanism for hypothesis testing would somehow lead to exploratory research (that is, re-

search that doesn't involve a priori hypotheses) being sidelined and seen as second-class. However this concern is illogical and speaks to a peculiar insecurity held by those who value exploration in science. All that Registered Reports do is to clarify the distinction between confirmatory hypothesis testing and more exploratory forms of analysis and hypothesis generation, both of which are valuable to science and welcomed as part of the format. But if the mere act of *distinguishing* one from the other denigrates exploratory research then what does that say about the value our community places in exploratory research in the first place? High rates of HARKing in psychology, as identified by Norbert Kerr and Leslie John (see chapters 1 and 2), show that exploratory research is indeed regularly crammed into an ill-fitting confirmatory framework, shoehorning Thomas Kuhn into Karl Popper for the sole purpose of achieving publication. Rather than criticizing Registered Reports for adding clarity to a system that champions obfuscation, why weren't the traditionalists developing parallel publishing initiatives that celebrate purely exploratory science? Indeed, if exploration mattered so much to them, why hadn't they done so years ago?[17]

- *Registered Reports can be gamed by "preregistering" a study that is already completed*. **Verdict: True only for fraudsters.** It's a curious sociological phenomenon that many otherwise reputable psychologists genuinely believe this is a possibility. Do they really hold their colleagues in such low regard? In any case, for Registered Reports such a strategy would not only be fruitless but is impossible without committing fraud. When authors submit a Stage 2 manuscript it must be accompanied by a laboratory log indicating the range of dates during which data collection took place together with a certification on behalf of all authors that no data (other than pilot data in the Stage 1 protocol) was collected *prior* to the date of IPA. Time-stamped raw data files generated by the preregistered study must also be deposited in a public archive, with the time-stamps postdating in-principle acceptance. Submitting a Stage 1 protocol for a study that had already been completed would therefore require complex falsification of laboratory records and data time stamps. Even putting aside the fact that such behavior

would be clearly fraudulent, traditionalists who raise this concern overlook a major problem with such a strategy: based on the comments of reviewers, editors usually require changes to the proposed experimental procedures following Stage 1 review. Even minor changes to a protocol would be impossible if the experiment had already been conducted, and would therefore defeat the purpose of preregistration. Unless the authors were willing to engage in further dishonesty about what their experimental procedures involved—a degree of fraud that is beyond redemption—"preregistering" a completed study would be a highly ineffective publication strategy.

- *Registered Reports won't stop fraud*. **Verdict: Straw man.** This was a common reaction and another straw man because Registered Reports are not designed to stop fraud. No publishing mechanism, least of all the status quo, can protect science against complex and premeditated acts of misconduct. What Registered Reports achieve, above all, is to eliminate publication bias along with the pressure to massage data, reinvent hypotheses, or behave dishonestly in the first place.
- *Registered Reports lock authors into publishing with a specific journal*. **Verdict: False.** Authors are free to withdraw their Registered Report submissions at any time—there is no binding contract with the journal. The only requirement is that study withdrawal after IPA leads to the publication of a Withdrawn Registration, which includes the abstract from the Stage 1 submission together with a reason for the withdrawal. This ensures that the process is transparent to the scientific community.
- *Registered Reports* fail *to lock authors into publishing with a specific journal*. **Verdict: Red herring.** Some traditionalists have criticized Registered Reports for exactly the opposite reason: that it could be gamed by authors precisely *because* there is no binding contract with the journal. Their argument goes like this. Suppose a researcher has a Stage 1 protocol accepted in principle with a specialist journal. They conduct their study but find something amazing and unexpected in the results that they feel could be sold to *Nature* or *Science*. What would stop them from withdrawing their paper from the specialist journal and resubmitting it as a conven-

tional (unregistered) article to a more prestigious outlet? The answer is: nothing. But there is a catch. Authors would do so knowing that their choice would be transparent to their peers because withdrawing a paper after IPA triggers publication of a Withdrawn Registration. It would interesting to see how an author's peers would react to a reason for withdrawing a Registered Report on the grounds that: "After finding something remarkable and unexpected, we decided we could publish this in a more prestigious journal." In many ways, such transparent careerism would be refreshing.

- *Registered Reports are not suitable for exploratory science or for developing new methods where there are no hypotheses*. **Verdict: Red herring.** This is a common objection but irrelevant because the format isn't designed to be applicable to anything other than hypothesis-driven science, which makes up a large bulk of published research in psychology and beyond.
- *Registered Reports are suitable only for one-shot experiments, not a series of sequential experiments where the outcomes of one experiment feed into the design of the next*. **Verdict: False.** At many of the adopting journals, authors can register experiments sequentially, each time adding to the previous set. At each stage in the cycle the previous version of the paper is accepted, eliminating the risk that the addition of later registered experiments could jeopardize publication of the earlier ones.
- *Reviewers of Stage 1 submissions could steal my ideas and scoop me*. **Verdict: Possible but highly unlikely.** Scooping is the bogeyman of science. Everyone knows someone who overheard someone talking with someone about a story in which someone got scooped. Maybe. In fact such cases are very rare.[18] Concerns about being scooped do not stop researchers applying for grant funding or presenting ideas at conferences, both of which involve releasing ideas to a much larger group of potential competitors than would typically see a Stage 1 Registered Report (which usually isn't published until the study is completed). It is also noteworthy that once in-principle acceptance is awarded, the journal cannot reject the Stage 2 submission because similar work was published elsewhere in the meantime. Therefore, even in the unlikely event of a re-

viewer rushing to complete a preregistered design ahead of the authors, such a strategy would bring little career advantage for the perpetrator, and would very possibly backfire.[19]

- *If Registered Reports were mandatory or universal they would . . .* **Verdict: Straw man and slippery slope fallacy.** However this sentence ends, the objection is irrelevant because we never proposed that Registered Reports should be mandatory or universal—indeed quite the opposite. The argument that Registered Reports should be a universal *option* for hypothesis-driven research is quite different to the argument (proposed by nobody) that it should be obligatory across all science.
- *A major previous discovery (e.g., mirror neurons) would never have been possible under a Registered Reports model, therefore Registered Reports would hold back science*. **Verdict: Arguably false but irrelevant even if true.** This concern is beside the point because Registered Reports have never been suggested as a replacement for exploratory science, only as an enhancement of hypothesis-driven science. But even putting that fact to one side, how can we be sure that discoveries such as mirror neurons wouldn't have emerged serendipitously from a Registered Report? A major misconception of Registered Reports is that they hinder serendipity when in fact they protect it. To illustrate this point, suppose you conducted a standard (unregistered) study and found something serendipitous that you believe is surprising and groundbreaking. What do you suppose will happen when you submit your findings to a journal through a conventional (unregistered) publishing route? Because the results are surprising the reviewers are likely to be skeptical about them, holding up publication for months or years while you argue your case or run additional experiments. You might even find that the barriers are too great and give up, dumping the results in a file drawer. Now suppose you conducted exactly the same study as a Registered Report. At Stage 2, reviewers can't recommend rejection on the basis of the results; therefore your serendipitous finding is protected.[20] This prompts us to turn the tables and ask: how many serendipitous results such as mirror neurons might have even been reported even *sooner* if they were revealed within Registered Reports?

- *We don't need Registered Reports because we have replication.* **Verdict: False.** This argument ignores the fact that direct replication in psychology is extremely rare and associated with many disincentives, not least of which is the contempt shown by journals for replication studies. Registered Reports, however, provide a perfect avenue for replications by provisionally accepting papers before authors invest the resources into conducting them. What better incentive could you have for persuading authors to consider conducting replication studies?
- *We don't need Registered Reports because protocols are already assessed through grant reviews.* **Verdict: False.** The first time I heard this criticism I couldn't believe the person was serious. Any psychological scientist who has reviewed or applied for a major grant knows that such applications contain nothing close to the level of technical detail that is required for a Stage 1 Registered Report. Grant applications propose the bigger picture of a planned program of work; they rarely drill down into the specific details of individual experiments to a degree that is required for a specific Stage 1 protocol. And even for the occasional cases where a protocol is sufficiently detailed, funded grant applications are almost never published so who would know whether the researcher did what they said they would?[21] A private preregistration that is never published is worthless to the scientific community.
- *With publication virtually guaranteed, authors of Registered Reports will conduct their experiments poorly, leading to meaningless results.* **Verdict: False.** This rather cynical objection emerges fairly regularly from traditionalists. As one put it: "If you're a young researcher and you get your good idea preaccepted based on the question and design, then it's just more efficient to do a quick, sloppy analysis and damn the results—after all, who cares? The paper was already accepted. Time to move on to the next one."[22] Aside from portraying early-career scientists as little more than ladder climbers, this argument ignores the fact that Stage 1 submissions must include outcome-neutral tests and quality checks for ensuring that the proposed methods are capable of testing the stated hypotheses. Stage 2 submissions that fail any critical

outcome-neutral tests can be rejected, providing an inherent safeguard against sloppy science. This objection also disregards the fact that Stage 1 review involves stringent and detailed assessment of the proposed analyses. It is no more possible for authors to conduct a "quick, sloppy analysis" as part of a Registered Report than for a standard conventional article—indeed it may be a lot less likely for a Registered Report.

- *The case for Registered Reports assumes that scientists act dishonestly, and sends the message that there is no trust in the scientific community.* **Verdict: Non sequitur and red herring.** This argument rests on the false premise that bad practice is synonymous with deliberate deceit. As we have seen, however, bias and questionable research practices can be unconscious or stem from ignorance without implying any dishonesty. At a deeper level, the objection misdirects us to place a greater emphasis on how psychological science is perceived externally, and how researchers feel, than on how the research is actually conducted. Regardless of whether bias and questionable practices are conscious or unconscious, the solutions are the same.
- *Registered Reports are based on a naive view of the scientific method.* **Verdict: False.** Registered Reports provide a way to protect the integrity of the deductive scientific method, but they do not elevate deductive science above alternative exploratory approaches. One might just as easily argue that a better drug for treating cancer is "naive" because it doesn't treat hepatitis. There is also a curious inconsistency inherent in this viewpoint. Some traditionalists may well believe that the hypothetico-deductive model is the wrong way to frame science, but if so, why do the very same researchers routinely publish articles that report *p* values and purport to test a priori hypotheses? Are they merely pretending to be deductive in order to get their papers published? Registered Reports ensure that when researchers are truly engaging in deductive science that they do it in as unbiased a way as possible and that they are rewarded appropriately for doing so. Those who criticize Registered Reports on these grounds are not actually arguing against Registered Reports. They are criticizing the fundamental

way research is taught and published in the life sciences, despite supporting that very system themselves and without proposing any alternative.

- *Registered Reports will overload the peer-review system*. **Verdict: Unknown but probably false.** It is true that Registered Reports involve two stages of peer review at the same journal, each of which is likely to involve at least one round of manuscript revision. However, this is offset by the fact that authors are much less likely to be successively rejected by multiple journals, as pointed out nicely by neuroscientist Molly Crockett in response to my *Cortex* open letter:

> [T]he value of this system is that a given manuscript will (ideally) only go through a single review process—so in terms of collective hours spent reviewing papers, your proposal may actually reduce the burden on the scientific community. Consider the process we have now. Papers often face a string of rejections before getting published (and often rejections are based on data, not methods—e.g., null findings). A given paper may go through the review process at 3 or 4 different journals before getting published—so anywhere from 6 to 12 (or more) reviewers may take the time to review the paper. This is extremely inefficient both for reviewers, and for authors, who must spend a substantial amount of time re-formatting the manuscript for different journals. None of this is time well spent. In contrast, the extra time involved for authors and reviewers in your proposed system *is* time well spent—the steps you outline guard against all sorts of problems that are rife in the scientific literature.[23]

- *Registered Reports will lead to researchers bombarding journals with protocols that have no funding or ethics and will never happen*. **Verdict: False.** As one critic said: "Pre-registration sets up a strong incentive to submit as many ideas/experiments as possible to as many high impact factor journals as possible."[24] Armed with IPA, the researcher could then prepare grant applications to support only the successful protocols, discarding the rejected ones. How-

> ever, this entire objection is beside the point because Stage 1 Registered Reports must include a statement confirming that all necessary support (e.g., funding, facilities) and approvals (e.g., ethics) are already in place and that the researchers could start promptly following IPA. Since these guarantees could not be made for unsupported proposals, the concern is moot.

As we can see, many of the critical reactions to Registered Reports were based on misunderstandings, logical fallacies, or ideological objections parading as rational counterarguments. In the wake of the *Guardian* letter we also faced a remarkable intensity of ad hominem attacks. Through various channels we were accused of being "self-righteous," "sanctimonious," "fascists," "a head prefect movement," "Nazis," "Stasi," "crusaders" on a "witch hunt," and worse. In one widely circulated e-mail, a professor who I happened to know went so far as to belittle the 80 scientists who signed our *Guardian* letter, stating: "Looking at the Chambers letter, I was struck by the lack of scientific weight of the signatories."[25] That the mere suggestion of a new type of article in science could provoke such aggression was telling about what the aggressors sought to protect. As Niccolò Machiavelli wrote over five centuries ago, "the innovator has for enemies all those who have done well under the old conditions."

Despite the fact that most of the critical reactions to Registered Reports were in my view deeply flawed, several points did have merit. One concern was that the time taken to review Stage 1 protocols could be incompatible with short-term student projects, where students would usually lack the time to wait for peer review and provisional acceptance before starting data collection. There are at least two possible solutions to this problem. The first is to accept that operating within such a rigid schedule is incompatible with Registered Reports and either conduct such studies as unregistered or instead preregister protocols for student projects in a database without peer review, such as the Open Science Framework. A more radical solution would be to reorganize undergraduate student projects into a daisy-chain system where students work for several months on a Stage 1 protocol while simultaneously implementing the provisionally accepted protocol from a previous year's student. Under this system, students would never implement the specific protocol that they submitted for peer review but they would nevertheless gain intensive training in all aspects of deductive science.

A second concern is that the sin of bean counting (see chapter 7) could make Registered Reports an unattractive choice for younger scientists. Because Registered Reports set rigorous methodological standards, requiring large sample sizes and higher statistical power, researchers can find that their experiments take longer to complete than they otherwise would under the status quo. As we saw in chapter 3, psychology and neuroscience are endemically underpowered, which permits researchers to publish a higher volume of lower-quality papers, rich in post hoc storytelling but making only a limited and biased contribution to knowledge. Registered Reports turn this equation upside down. Researchers who publish Registered Reports are likely to publish a fewer number of larger and more credible papers; however, as long as the community values quantity over quality then a more credible publication record is not guaranteed to provide young scientists with more secure careers in science—and this is a problem that can be fixed only by senior scientists changing the way they assess junior researchers for jobs, grants, and career progression. A related concern is the conservatism of the most prestigious journals, which despite claiming to publish the highest-quality research nevertheless rely on publication bias to select which papers to accept. To compete in the current academic system, young scientists need to be strategic in where they send their work, so if Registered Reports were offered only within specialist journals their reach and appeal would hit a glass ceiling. In the shorter term, the solution to this problem is the central goal of the Registered Reports initiative: to see the format launched within *all* journals that publish hypothesis-driven science, regardless of prestige. In the longer term, the solution—as we will discuss later—is to do away with journals altogether, rejecting the premise of "prestigious" outlets and allowing the quality and contribution of individual studies to speak for themselves.

A third limitation of Registered Reports is that it is unclear to what extent it can be applied to analyses of preexisting data. We saw in chapter 4 how analysis of existing data archives can be used to answer important questions that may not occur to the investigators who conduct the original studies. This raises the question of whether analysis of preexisting data could be preregistered under a Registered Reports mechanism without the process being biased by the authors having prior knowledge of the outcomes. A potential solution to this problem—at least for sufficiently "big" data—is to consider such existing data sets as split-half discovery and replication

samples. The system would be analogous to a card game in which one player deals while the other cuts the deck: After vowing on the record that they had never viewed the dataset in question, authors would submit a proposed analysis to the journal. If the proposal passes prestudy review and is provisionally accepted, then the journal would then decide (using a random algorithm) which half of the data is the discovery sample and which is the replication sample—that is, a random cutting of the deck. Under this system, a registered secondary analysis would be considered to produce a finding of note only if it replicated in each subsample.

Finally, one important question is whether there is any evidence that Registered Reports will be effective in reducing publication bias and questionable research practices. As with any new initiative, prior evidence of effectiveness cannot and does not (yet) exist.[26] From a logical point of view, barring failure of the peer-review process, prespecification of hypotheses and analysis plans is guaranteed to eliminate the practices of *p*-hacking and HARKING. Similarly, unless Daryl Bem was right all along about the existence of precognition, accepting papers before results are known renders them immune to publication bias. But whether eliminating these practices will lead to more reproducible science is an empirical question that has yet to be answered. In 2014, the *International Journal of Radiation Oncology, Biology, Physics* launched a randomized trial of Registered Reports, results of which are pending. Among other outcome measures, the editors are testing whether the rate of positive, negative, and indeterminate results will differ between studies that are assigned randomly to either a standard unregistered format or to a Registered Report.[27] There are also signs in the clinical trials literature that preregistration, in general, may reduce publication bias and/or researcher bias. A 2015 analysis of medical trials in the prevention of heart disease found that, since the advent of mandatory clinical trial registration in 2000, the percentage of published trials showing a statistically significant effect dropped from 57 percent (17 out of 30) to just 8 percent (2 out of 25).[28] Preregistration might therefore help put the brakes on false discoveries.

Three years after the launch of Registered Reports the tone of the discussion has shifted. The storm of personal attacks has subsided, and the opposition from many traditionalists appears to have softened. The debate is far from over, but important progress has been made. Some who initially resisted the initiative have become supporters, and we are seeing completed

examples of Registered Reports now appearing in the literature.[29] At the time of writing, more than 40 journals have adopted the format, and not just in psychology and neuroscience but also in cancer biology, empirical accounting, nutrition research, political science, and psychiatry. In addition, by the time this book is in print, a number of "high-impact" journals are likely to be offering them, including *Nature Human Behaviour*.[30] The initiative has been heralded by the UK Academy of Medical Sciences as one of several promising solutions to improving research transparency and eradicating publication bias.[31] In parallel, more than 750 journals across the full range of sciences have agreed to review their adoption of open science as part of the Transparency and Openness Promotion (TOP) guidelines, a process that involves the consideration of Registered Reports.[32] In March 2014 we also established the Registered Reports Committee at the Center for Open Science. This committee, which I currently chair, aims to develop and promote Registered Reports as a new way to improve the credibility of published research.

Perhaps the most significant step forward for Registered Reports came in November 2015. On the 350th anniversary of launching the world's first scientific journal, the Royal Society officially launched the initiative within its journal *Royal Society Open Science*. This is a major development not only because it is the first endorsement of Registered Reports by a learned society, but because it establishes the format well beyond psychology to cover the full spectrum of more than 200 physical and life sciences.[33] From here a future beckons in which Registered Reports may become a popular format for science in general, offered at all journals that publish the outcomes of hypothesis testing. If our goal is to establish a reproducible knowledge base, then the literature must include at least some papers that are published based on theoretical value and methodological rigor, independently of results.

Preregistration without Peer Review

Registered Reports are rapidly becoming available across a range of scientific fields including psychology; however in many cases journal-based preregistration may not be possible for researchers. The authors' preferred journal may not yet offer the format, or, as with student projects discussed earlier, the authors may need to preregister and commence data collection

immediately, with no time to go through the Stage 1 review process. Even when Registered Reports are not an option, it is still possible to preregister studies in one of several available online registries.[34] Protocols deposited in registries are not peer reviewed, but they offer authors the opportunity to prespecify their hypotheses and analysis plans, which can help assure skeptical reviewers that a particular hypothesis or analysis was preplanned rather than post hoc. As psychologist Leif Nelson puts it: "In a world of transparent reporting, I choose preregistration as a way to selfishly show off that I predicted the outcome of my study."[35]

Unreviewed preregistration has pros and cons compared with Registered Reports. Because there is no peer review, it has the advantage of allowing researchers to proceed sooner with data collection. On the other hand, the lack of review means that a protocol published in a registry has no assurance of leading to a peer-reviewed publication. Furthermore, unreviewed preregistration is vulnerable to the criticism that it allows authors the opportunity to publish vague protocols that are not replicable and that permit all the usual researcher degrees of freedom. Therefore, peer-reviewed articles eventually arising from unreviewed protocols could still contain *p*-hacking or other forms of selective reporting. Evidence from the clinical trials literature suggests that unreviewed preregistration is indeed prone to bias. A 2009 analysis found that of 147 adequately preregistered medical trials, 31 percent later changed their primary outcome measures, and 13.9 percent were even "preregistered" *after* the completion of the study.[36] A more recent analysis of 89,204 trials found that 31.7 percent had changed their primary outcome measures.[37] The same team later found that, consistent with *p*-hacking, alteration of primary outcome measures was associated with the reporting of statistically significant outcomes.[38] In the wake of these findings, Ben Goldacre and his team established the COMPARE project to monitor the extent to which medical researchers adhere to their preregistered protocols.[39] The results so far have not been encouraging: at the time of writing, 58 of 67 trials published since 2014 have either failed to report all preregistered outcome measures or have silently added new "primary" outcomes that were not preregistered.

Although preregistration is uncommon in psychology, limited evidence points to similar problems with unreviewed protocols. In 2015, Annie Franco and colleagues compared the preregistered protocols with final published articles for 32 studies entered in the Time-sharing Experiments for

the Social Sciences registry, an initiative of the National Science Foundation.[40] They found that 41 percent of published articles reported fewer experimental conditions than in the associated protocol and that 72 percent reported fewer outcome measures. Consistent with selective reporting practices, including *p*-hacking and HARKing, effect sizes for outcomes that were reported were consistently larger than for those that went unreported.

The warning for psychology here is clear. While unreviewed preregistration offers greater protection against bias than no preregistration, the lack peer review of the protocol—and the lack of comparison between the preregistered protocol and the final report—means that the credibility of preregistered studies will vary significantly. Unreviewed preregistration is therefore likely to be less robust than Registered Reports, where peer review and continuity between protocol and outcome is inbuilt.[41]

Solving the Sin of Unreliability

How can we make psychological science more robust and reproducible? In chapter 3 we considered in detail the sin of unreliability, caused by indifference (and in some cases hostility) toward replication, lack of statistical power, publication of vague and nonreplicable methods sections, misuse and misapprehension of frequentist statistical methods, and failure to retract clearly unreliable papers from the scientific record. We also discussed some of the possible solutions, including a greater emphasis on replication and high statistical power, adoption of Bayesian inferential methods to allow more meaningful hypothesis testing, increased adversarial collaborations to overcome ego-driven biases, and the implementation of higher methodological reporting standards at journals. Registered Reports have the potential to provide a natural antidote to many of these problems. Because they are reviewed and accepted in advance of data collection, authors can rest assured that their efforts in replicating a previous study will not end up in the file drawer. Registered Reports improve reliability by setting minimum a priori power requirements and welcoming Bayesian inferential methods. However, despite the great promise held by Registered Reports, the format is currently too embryonic to provide a one-size-fits-all solution for unreliability; moreover because it focuses on publishing (scientific outputs) it may never provide an all-encompassing solution that addresses the

full range of the scientific process, including exploratory research and inputs from funders. This raises the question of what other initiatives could be instituted, both in psychology and in wider science, to incentivize robustness and boost reliability. Here are some hypothetical initiatives that do not yet exist but that could provide additional advantages:

Development of a Reproducibility Index. As noted in chapter 7, science is knee deep in citation metrics that provide only limited value. One potentially useful metric, however, may be the extent to which a previously published finding has been replicated. The idea for a Reproducibility Index was suggested in 2013 by Adam Marcus and Ivan Oransky, who proposed it as a complement to the journal impact factor: "Rather than rate journals on how often their articles are cited by other researchers, let's grade them on how well those papers stand the most important test of science: namely, does the work stand up to scrutiny?"[42] There are several ways such a metric could be developed. One strategy would be to continuously monitor all literature following the publication of a key article, with the paper's score regularly updated either positively if the central finding is replicated, negatively if the central finding fails to replicate, or neutrally if no replication attempts are published.[43] The Reproducibility Index would thus provide a running gauge of the known reliability of particular findings and papers. If such a measure were rolled out widely across psychology, indexes at the level of specific results could be clustered to indicate the evidential reliability of subdisciplines, journals, and institutions, though as always the reliance of any one metric would not be recommended.

The greatest barrier in developing a Reproducibility Index would be the initial workload. Establishing the foundation of such an initiative would require a careful retrospective analysis of most (if not all) of the psychological literature, which amounts to hundreds of thousands, if not millions, of published papers. Since the definition of what constitutes a "replication" is in many cases subjective, it would be impossible to fully automate such a project[44]—it would instead require the coordinated action of a large number of experts within different fields to chart the "replication profile" of the entire evidence base. Starting from the oldest paper, each individual assessment would be based on an automated forward-search of all peer-reviewed literature that either cited or otherwise undertook a sufficiently similar methodology to the original study. With the original study and a manage-

able subset of subsequent papers in hand, the assessor would then judge whether (a) each of the newer studies performed a replication of the original study, either direct (close) or indirect/conceptual—with the nature of any variation noted; and (b) the extent to which that replication, if present, succeeded. Successful and unsuccessful replications could lead to the original study being scored positively or negatively accordingly, with greater scores applied for closer replications and greater degrees of success or failure. To maximize reliability, many different experts would ideally assess each paper, with the overall score determined by an aggregate analysis of the distribution of individual scores. Once the profile of the current literature was in place, the project could then shift into a rolling mode, scanning new entries to the literature as they appear. By taking advantage of machine-readable tags in published papers, it is possible that rolling updating of the Reproducibility Index could be at least partially automated.

Of course, the key question underlining such a project is who would perform the daunting task of establishing such a mechanism? There is no easy answer to this question, but the solution—if there is one—must lie in crowd sourcing. According to records held by the International Union of Psychological Science, it is estimated that there are approximately one million psychology researchers worldwide, represented by 87 national professional bodies.[45] By distributing the workload across many thousands of members of this population (much like peer review) we could make the project feasible if additional funding could be attracted to provide centralized administrative support.[46] Achieving such a project would nevertheless require an unprecedented unity of purpose within the psychological community.

Pottery Barn rule by journals and funders. One innovative solution to unreliability discussed in chapter 3 was Sanjay Srivastava's Pottery Barn rule in which journals would commit to publishing replication attempts of novel findings that they previously published. This idea could also be extended to include funding agencies, with a percentage of total science spending ring fenced to support replication attempts of previously funded projects. Although this would in the short term reduce the net funds available for original research, in the medium term it could improve reproducibility and also bring significant cost savings. Across the whole of biomedical sci-

ence, it has been estimated that lack of reproducibility costs the United States over $25 billion annually.[47] Diverting a relatively small amount of public money to support reproducibility and replication early in the scientific process has the potential to reap significant benefits downstream, both into terms of discovery and economic efficiency.

Regular multicenter replication initiatives. In any field where direct replications are only rarely undertaken, it is vital that there are regular group-led initiatives to monitor reproducibility. Brian Nosek and the Center for Open Science have led the way in championing such approaches in psychological science. In chapter 3 we covered the 2014 special issue in the journal *Social Psychology* that reported replication attempts across a wide range of studies, covering recent influential papers to seminal work of the 1950s. In 2015, Nosek and a team of 270 researchers followed up with an even larger reproducibility project, this time attempting to replicate 100 major studies. They found that just over 1 in 3 prior results could be reliably reproduced.[48] Until replication and minimally biased practices are standard in psychology, such initiatives will be crucial for assessing the credibility of the evidence base. They must, therefore, continue to be supported and expanded to cover more subfields.

Registered Reports funding schemes. To date, Registered Reports have been offered only by journals, with authors required to have all necessary funding and approvals in place before they submit a protocol. However, it is possible to perform protocol review even sooner, integrating the review of protocols by journals and funders to maximize efficiency and impact. According to the proposed system, authors would submit their research proposal before they have funding.[49] Following simultaneous review by the both the funder and the journal, the strongest proposals would be offered financial support by the funder and in-principle acceptance for publication by the journal. This model carries a range of potential benefits for all stakeholders. The journal stands to benefit by publishing work that is both cutting edge and reproducible—criteria that are difficult to achieve in single papers published through conventional means. At the same time, the funder stands to benefit by supporting a set of projects that are guaranteed to be published in a respected journal, eliminating publication bias

and maximizing transparency. This ensures that the specific work supported by the funder is undertaken; and there may also be administrative efficiencies for the funder where the review process is managed, at least in part, by the journal (with the funder of course maintaining a key role in deciding which work is supported). Authors also stand to benefit by having their publications accepted in a respected journal before they start their research, and through a mechanism that not only minimizes researcher bias but also eliminates the incentive for authors to engage in biased research practices in the first place. The papers they produce from such a mechanism are therefore likely to be impactful and well cited. Finally, the scientific community, more widely, stands to benefit from increasing the stock of research that is timely, important, transparent, and reproducible.

Solving the Sin of Data Hoarding

In chapter 4 we witnessed the widespread lack of data sharing in psychological science. Despite clear advantages for science as a whole, few researchers deposit their data in public repositories, and most even fail to share data with other researchers when requested to do so. Central among the emerging reforms are the Peer Reviewers' Openness (PRO) initiative and the Transparency and Openness Promotion (TOP) guidelines. These initiatives are very different but highly complementary. PRO mobilizes the democratic power of peer reviewers to insist that archiving of data and materials, or a publicly stated reason for not archiving, is a critical prerequisite for the peer to provide an in-depth review. The TOP guidelines, on the other hand, seek to encourage greater transparency at the level of policy by asking journals and organizational signatories to evaluate and publish their adherence to various transparency standards and levels of adoption. Although the TOP guidelines do not require journals to adopt open practices, the fact that their degree of compliance will be published provides an incentive to adopt as high a level of transparency as possible.

Both the PRO and TOP initiatives are tremendously promising because they seek to ensure visibility and accountability of open science practices. But they are not enough. For all publicly funded psychology research, it

TOP Guidelines Adopted by More Than 750 Journals and 60 Organizations

	Level 0	Level 1	Level 2	Level 3
Citation Standards	Journal encourages citation of data, code, and materials, or says nothing.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article provides appropriate citation for data and materials used consistent with journal's author guidelines.	Article is not published until providing appropriate citation for data and materials following journal's author guidelines.
Data Transparency	Journal encourages data sharing, or says nothing.	Article states whether data are available, and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Analytic Methods (Code) Transparency	Journal encourages code sharing, or says nothing.	Article states whether code is available, and, if so, where to access it.	Code must be posted to a trusted repository. Exceptions must be identified at article submission.	Code must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Research Materials Transparency	Journal encourages materials sharing, or says nothing.	Article states whether materials are available, and, if so, where to access them.	Materials must be posted to a trusted repository. Exceptions must be identified at article submission.	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Design and Analysis Transparency	Journal encourages design and analysis transparency, or says nothing.	Journal articulates design transparency standards.	Journal requires adherence to design transparency standards for review and publication.	Journal requires and enforces adherence to design transparency standards for review and publication.
Study Preregistration	Journal says nothing.	Journal encourages preregistration of studies and provides link in article to preregistration if it exists.	Journal encourages preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
Analysis Plan Preregistration	Journal says nothing.	Journal encourages preanalysis plans and provides link in article to registered analysis plan if it exists.	Journal encourages preanalysis plans and provides link in article and certification of meeting registered analysis plan badge requirements.	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.

TOP Guidelines Adopted by More Than 750 Journals and 60 Organizations (*cont.*)

	Level 0	Level 1	Level 2	Level 3
Replication	Journal discourages submission of replication studies, or says nothing.	Journal encourages submission of replication studies.	Journal encourages submission of replication studies and conducts results blind review.	Journal uses Registered Reports as a submission option for replication studies with peer review prior to observing the study outcomes.

At the time of writing, the TOP guidelines have been adopted by more than 750 journals and 60 organizations. For their policy on data and materials to also be compliant with the PRO initiative, the journal or organization would need to achieve Level 2 or above for the *Data Transparency* and *Research Materials Transparency* standards. Table adapted from Nosek et al. (2015).

is in the public interest for archiving of study materials, analysis code, and anonymized data to be *mandatory*, with the mandate enforced by universities and funders, and possibly through legislation. Of course, in some cases it may not be possible to allow full public access to such archives; some materials may be copyrighted or subject to other legal restrictions, and data might be potentially attributed to specific individuals even when the researcher believes it is sufficiently anonymized. In balancing the ethical risks of sharing data, researchers can conduct a risk assessment that takes into account two factors: the sensitivity of the data acquired and the estimated likelihood of anonymization being thwarted to reveal a participant's identity. In special cases, researchers may also have good reasons to apply a time-limited embargo on data access, on the grounds that they plan to publish several papers arising from a single, indivisible data set. Such considerations should rightly influence the freedom of access to archived data—questions of *who* can access it and *when*. However it is unacceptable that such considerations determine whether data is archived in the first place. All data should be archived because, sooner or later, data that are not archived are lost to science and to future generations of scientists. The vast majority of everyday studies in cognitive, experimental, and social psychology could immediately release simple anonymized data with minimal to no ethical concerns, and for those where risks are higher, models of gated access—either temporarily or in the long term—can be applied.

Solving the Sin of Corruptibility

In chapter 5 we journeyed beyond the common gray area of questionable research practices to study the outer reaches of fraud. What we find is a field that is powerless at preventing fraud and incapable of protecting whistle-blowers. Of all the deadly sins, the sin of corruptibility is the least forgivable because it perpetrates such grave injustices against the most vulnerable researchers in our community. It is also the most challenging to reform because many of the solutions will require coordination with external organizations and lawmakers.

Random data audits. Acts of fabrication often leave a statistical footprint in the data, as we saw with the cases exposed by Uri Simonsohn and alleged by the Förster whistle-blower. However, the exposure of such fraud cases should not depend on self-regulation by lone "data detectives." What psychology (and science more widely) needs is an external regulator to systematically audit research data. With additional investment of resources, this monitoring could be accomplished by existing mechanisms, including the UK Research Integrity Office and the US Office of Research Integrity. Monitoring could be dovetailed with data archiving: automated algorithms could search available data in published articles and archived repositories, either systematically or randomly depending on the breadth and depth of the analyses, and with the potential for additional human auditing in cases flagged as potentially suspicious. Like drug testing in sport, rigorous data audits conducted in accordance with due process (including the presumption of innocence) could be a powerful mechanism for detecting and deterring acts of fraud.

Profiling of researchers. Could personality profiling of researchers help prevent cases of fraud before they happen? In 2010, psychologists Kevin Williams and colleagues from the University of British Columbia reported that, of several personality measures, plagiarism in psychology undergraduates was correlated most strongly with psychopathy.[50] Assuming that this relationship is causal and extends to data fabrication, this raises the question of whether students and staff who score high on psychopathy may be more likely to commit academic fraud. Should such individuals even be weeded out of the system, as Kate suggested following her ordeal (see

chapter 5)? Although prevention is innately appealing, such an approach raises grave ethical and practical concerns. Williams and colleagues explain why:

> It seems unlikely that school boards and university senates would approve of mass prescreening of students for psychopathy. Any attempt to determine probability-of-expulsion in advance suggests an unsavory "guilty until proven innocent" approach. . . . Even if prescreening were to be approved, there is no established cutoff score for psychopathy in nonoffender populations. . . . Even if scores were kept confidential, labeling could be extremely harmful to the student. The surveillance of high scoring individuals would be highly problematic ethically and practically.[51]

Even though profiling is premature and arguably quite sinister, there is clearly a need for more research on the personality characteristics of academic fraudsters. At the same time, selection panels for PhD programs and academic appointments should favor applicants with a proven track record in honesty, rigor, and openness—characteristics that are likely to provide a safeguard against fraud.

Criminalization of academic fraud. One of the most surprising facts about academic fraud is that it is seldom prosecuted as a criminal offence. While perpetrators face dismissal from their jobs or student candidature, and their academic careers are usually terminated, they are unlikely to face fraud charges. Diederik Stapel's case stands as an exception to this rule—recall from chapter 5 that he avoided trial in the Netherlands by plea-bargaining to 120 hours of community service, albeit a sentence usually administered for minor offences. The prosecutor's office concluded that Stapel had "acted against scientific integrity" while also ruling, oddly, that he had "not misused public funds, because they were given for doing research that was performed," even though the research was fraudulent.[52] Had Stapel embezzled funds for direct personal gain, he would no doubt have faced more serious criminal charges.

Is it acceptable that academic fraud is treated at most as a minor offence, if an offense at all? Given how easy it is to get away with data fabrication, the career advantages it brings, the public resources its wastes, and the collateral damage it causes to students and colleagues, there is an argument

that academic fraud should be added to the criminal statutes and prosecuted as a more substantial offence. In 2014, such a mechanism was put to the test in Australia in the case of Parkinson's disease researchers Bruce Murdoch and Caroline Barwood from the University of Queensland. Following an internal investigation triggered by a whistle-blower, Murdoch and Barwood resigned, and two of their coauthored papers were retracted under allegations that the studies never took place. The university then referred the case to the Queensland Crime and Corruption Commission, which, following its own investigation, charged Murdoch with three counts of fraud, seven counts of fraudulent falsification of records, and five counts of general dishonesty.[53] Barwood was also charged with six counts of fraud or attempted fraud.[54] In March 2016, Murdoch pleaded guilty to all charges and was handed a two-year suspended jail sentence, with Barwood's trial still pending. This case may provide a basis for establishing similar legal mechanisms in other countries.

Protection and support of whistle-blowers. Our academic system fails abysmally to protect whistle-blowers. In the case of Diederik Stapel, the three junior researchers who exposed him would almost certainly have had more successful academic careers had they remained silent. And in Kate's case, the fact that she assisted and advised junior whistle-blowers led to her being sacked by a vengeful line manager, followed by ten years of academic exile. What can we do to better protect those who put their careers on the line to expose misconduct? The first principle must be to ensure that speaking out is as minimally disadvantageous as possible for whistle-blowers.[55] Institutions should adopt research integrity policies that commit them to supporting the careers of whistle-blowers. For a postdoctoral researcher like Kate, this could be extending a current salary so as to provide researchers with the opportunity to continue their work, either independently or under new line management. For a PhD student it could mean offering extended candidature and the chance to make up lost ground with a new project and supervisor. It must also be enshrined in law that whistle-blowing presents no barrier to promotion or retention. Finally, we must ensure that institutions have in place independent and transparent mechanisms for dealing with allegations of misconduct because it is rarely clear who whistle-blowers can safely approach with their concerns.

Replication. Direct replication by independent researchers is the ultimate method for rooting out error in science, whether caused by fraud, bias, or other mistakes.[56] Nonreplications alone should never be considered diagnostic of fraud, but they are the best way to ensure that the scientific record self-corrects in the long run. A culture in which independent replication of key findings is routine would also reduce the incentive to commit fraud in the first place, because it would ensure that no single finding is heralded as a "discovery" until it is has been independently repeated.

Solving the Sin of Internment

In chapter 6 we saw how much psychological science is "published" behind paywalls that make it inaccessible to the public, including nonacademic users. Of all the deadly sins, this is perhaps the easiest to remedy through individual actions and the growing momentum toward open access (OA) publishing by governments and funders. One simple action that all researchers can take is to ensure that they publish in journals designated as "green" by the Rights MEtadata for Open archiving (RoMEO) project.[57] RoMEO green journals allow posting of papers on a personal website or freely accessible repository immediately after acceptance (i.e., without an embargo). In the medium term, funders, institutions, and governments should further advance policies that not only mandate the publication of publicly funded research through OA routes, but which do so via fully OA journals rather than hybrid OA.

In the longer term, the debate surrounding OA publishing gives us pause to consider whether the current journal-based system is fit for purpose. Is it right that corporate publishers generate billions in profits on the back of free labor provided by scientists? Given that scientists perform all the essential work—the research, the peer review, and the editing—why do we even need publishers? The litmus test for any system facing obsolescence is whether we would invent it, as it is now, if didn't already exist. In their 2013 article (see chapter 7), Björn Brembs, Kate Button, and Marcus Munafò call for a radical reinvention of the academic publishing system to place it under the administration of university libraries, dropping the need for journals and supporting the totality of the peer-reviewed literature at a fraction of the current cost:

> We therefore would favor bringing scholarly communication back to the research institutions in an archival publication system in which both software, raw data and their text descriptions are archived and made accessible, after peer-review. . . . This reputation system would be subjected to the same standards of scientific scrutiny as are commonly applied to all scientific matters and evolve to minimize gaming and maximize the alignment of researchers' interests with those of science.[58]

Neuroscientists Sam Schwarzkopf and Dorothy Bishop have proposed similar ideas—a new platform for peer review and publishing that is divorced from journals.[59] Under Schwarzkopf's proposed system, authors submit their manuscript to a centralized review platform where it is considered by editors and reviewers before publication. However, unlike most journals, the editorial decision is based on scientific rigor rather than the usual criteria of novelty or impact. Schwarzkopf's system also opens up the possibility of different types of review, including not only the traditional review of the entire paper but also various forms of partial review including *rating-only reviews* where a larger number of reviewers score a paper without providing any written feedback, *specialist statistical review* where the reviewer focuses solely on the statistical validity of the analyses, and *data review* where reviewers are tasked with checking the authors' results by repeating their data analyses using the supplied analysis scripts. Following favorable reviews—and revision where necessary—the article, data, and materials would be published under a Creative Commons Attribution license (CC-BY), and each item would be separately citable to enable transparent reuse and assignment of credit. To assure maximum transparency, the peer reviews themselves would also be published alongside the final article, which could either be signed or anonymized.

Both Schwarzkopf and Bishop propose a model in which study protocols could be reviewed prior to results being gathered, in the same way that journals currently consider Stage 1 Registered Reports. Under Bishop's proposal, the protocol could be submitted before funding has been acquired, and even before a sufficient team has been assembled to conduct the research. Publication of the protocol could then be used to attract collaborators, similar to the Registered Replication Reports initiative offered by *Perspectives on Psychological Science*. Following peer review, protocol ac-

ceptance, and the formation of a suitable team, the published protocol could then be used as a basis for a funding application. This process may be attractive to funders because a large part of the peer review (which the funder would normally coordinate) has already been undertaken. The system would also be open to various other kinds of specialized article formats, including review articles, exploratory (non-hypothesis-driven) empirical articles, and methods development.

Once a paper is formally accepted on the platform, this new system turns the tables on traditional publishing. In addition to gating publication through prepublication review, the platform would also invite postpublication review, where other scientists are able to comment on and rate the accepted article. Those articles that are rated highly or that generate the most traffic would rise to the surface and become more prominent on the platform. Then, rather than authors submitting their work to zeitgeist journals such as *Nature* or *Science*, these journals (and others) could instead *bid* for articles on the platform. For example, they might invite authors to submit a shorter version for their journal that emphasizes wider implications (linked to the full version of the paper); alternatively, specialist science writers based within journals could publish overviews of the most popular or highest-rated papers.

By separating peer review and publishing from journals, this system has numerous advantages over traditional scholarly publishing. As Brembs and colleagues point out, it would be much cheaper than maintaining expensive journal subscriptions and OA publishing costs and would thus permit the redirection of university library budgets toward supporting the coordination and administration of the platform. The platform would also ensure that publication of academic research is based on the theoretical importance and methodological rigor of a study, with factors such as novelty and impact relevant only in determining the newsworthiness to secondary users *after* publication. And finally, it would forever solve the sin of internment by ensuring that fully OA publishing is an intrinsic product of the research process.

Solving the Sin of Bean Counting

In chapter 7 we considered the various ways in which psychology is failed by an academic system that idolizes superficial citation metrics, treats re-

search grants as outputs, and remains in the grip of an archaic system of academic authorship where assumptions about individual contributions are based on the crudest of measures: authorship rank.

A potential solution to the problem of authorship rank would be to replace it with a more transparent contributorship model that identifies the specific involvement in the research made by the different authors.[60] Under a contributorship model, for every publication, authors would be listed in alphabetical order, with authors listing their percentage contribution to each part of the paper.[61] Potential categories of contribution could include: theory and background, study design, data acquisition, provision of analytic tools/reagents, data analysis, interpretation of results, and manuscript preparation. Percentages would be allocated in two ways: relative to the sum of work completed by all authors on the paper, and relative to the totality of each individual's contribution. Each of these statistics would provide useful information to different users. On the one hand, the percent contribution relative to the paper would tell the community which authors contributed most to different aspects of a research project. If we take an example of a cognitive neuroscience paper led by a postdoctoral researcher and principal investigator, together with advice offered by a statistician, it might be the case that the postdoc did 100 percent of the data collection, and that the postdoc and statistician divided up the data analysis between them 80 percent to 20 percent. The principal investigator, however, might have had no involvement in the data acquisition or analysis but might have had the most substantial contribution to the theory and background and the study design, dividing this 60 percent to 40 percent with the postdoc. To comply with established conditions for authorship, all authors would also be expected to make a contribution to manuscript preparation.

On the other hand, the percent contribution to different domains relative to the author's total contribution could provide useful information for assessors in constructing profiles of individual researchers. By averaging contributions across publications within the different categories, a researcher might develop a profile as an all-rounder who takes a hand in all aspects of their research, or as an implementer who conducts studies and interprets the results but plays less of a role in theory and study design. Alternatively a researcher might emerge as an analyst who specializes in statistical methods, or as a theoretician and designer—as would be typical of many princi-

pal investigators. Profiles are also likely to change over time: as researchers become more senior and progress from being postdocs to principal investigators, their profiles may also switch from being all-rounders or implementers to being theoreticians and designers.

Replacing traditional authorship with a contributorship model has the advantage of achieving maximum transparency, but as with all systems that generate quantitative outputs, it is vital that such information is not overinterpreted. One risk of a contributorship model is the community drawing a false sense of precision from the percent contributions, which are likely to be uncertain in many cases.[62] This presents a significant danger for individual profiling, where, for example, funding agencies or university departments might decide to use contributorship profiles to triage applicants for grants or jobs. While this is a genuine threat, clinging to the traditional authorship model does little to protect us from such bean counting, as job applicants are already routinely triaged according to their quantity of first-authored publications or grants.

In the long run, to solve the sin of bean counting we must answer a fundamental question: how do we judge the quality of research and researchers? There is no metric, no simple heuristic, no set of calculations or algorithms. Perhaps we can take a cue from the physical sciences, which have long embraced subjective quality assessments by independent experts. This unattributed quote provided by pharmacologist David Colquhoun shows how one chemistry department assesses research excellence:

> The greatness of a faculty member is not judged simply by the members of the Department but rather by letters we collected, typically 10 to 15, from experts outside the Department, nationally and internationally. The question we ask of these experts is whether the research of the candidate has changed the community's view of the nature of chemistry in a positive way. It is not based on how much funds the candidate had brought to the University in the form of grants. It is not based on the number of published papers. It is not based some elaborate algorithm that weighs publications in journals according to the impact factor of the journal. It is based simply on establishing new knowledge. As a Department we do not discuss *h*-index metrics and we do not count publications or rank them as to who is first author. We just ask, has the candidate really changed significantly how we understand chemistry.[63]

Concrete Steps for Reform

Having reviewed the seven deadly sins and a range of possible solutions, what can each of us to do to safeguard the future of psychological science? What follows are series of suggested actions that we can take as researchers, organizations, and nonacademic citizens.

For junior and senior researchers:

1. Be aware of the dangers of confirmation bias, low statistical power, and *p*-hacking. For studies with clear a priori hypotheses, submit your work as a Registered Report or preregister it on an available registry such as Open Science Framework.
2. Be aware of hindsight bias and Hypothesizing After Results are Known (HARKing). For work that is truly exploratory, don't write up the paper so that it presents an a priori hypothesis in the introduction.
3. When writing the method section of an article, declare your degree of transparency by stating the "21 word solution" by proposed by Simmons, Nelson, and Simonsohn (2012): "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study."[64]
4. Sign the Peer Reviewers' Openness initiative.[65]
5. Adopt Bayesian hypothesis testing. Become familiar with the interpretative limitations of *p* values and conventional null hypothesis significance testing.
6. Avoid the trap of assuming that every new result published in the psychological literature is a true positive and a definite fact. Science doesn't proceed one fact at a time; it is more like the process of reclaiming land. If a particular finding of interest hasn't been directly and *independently* replicated then be cautious about extending it and applying it in ways that depend on it being credible.
7. Publish your research in RoMEO green or blue journals, prioritizing outlets that allow immediate (embargo-free) archiving of accepted papers. Never publish in journals that require subscription access and that don't permit archiving.[66]

For senior researchers:

1. Participate in joint replication initiatives in your field for the common good.
2. Implement policies within your laboratory, department, or institute for standardized publicly archiving of data, code, and materials.
3. Establish a "data partner scheme" with other labs for mutual checking of data analyses and curation strategies.[67]
4. Take signs of research misconduct seriously. Assume responsibility for dealing with fraud before it becomes someone else's problem.
5. Avoid judging job applicants on the basis of journal impact factors, grant income, quantity of first-authored publications, or any other superficial metric. Instead, ask applicants and their referees to describe their best papers, their contribution to the work, and what the research contributed to theory or practice. Value any proven commitment to open science practices and principles.
6. Sign the San Francisco Declaration on Research Assessment (DORA) as an individual.[68]
7. Above all, remember that junior scientists will learn by your example. If you indulge in the deadly sins, your protégés will too. And, in time, so will theirs.

For journals, funders, learned societies, or universities:

1. Sign the Transparency and Openness Promotion guidelines.[69]
2. For journals: If your journal publishes empirical papers that report the outcomes of hypothesis testing then offer Registered Reports as a new article option.[70]
3. For journals: Offer badges for open practices and require authors to complete methodological checklists at submission.[71]
4. For journals and funders: Introduce a policy for sharing of data and study materials that complies with the Peer Reviewers Openness' Initiative.
5. For universities and funders: Never judge the track record of applicants based on their record of grant income. Income should

be either ignored or balanced against the contribution of the published outputs to theory or practice.
6. Sign DORA as an organization.

For journalists and citizens:

1. For journalists and readers: Adopt a critical mindset when writing or reading news stories reporting the latest psychological science. Has the study been replicated, either independently or by the authors? Was the sample size large enough to justify the conclusions that are being drawn? What measures were put in place to counteract sources of research bias? Remember that extraordinary claims require extraordinary evidence.
2. For journalists: Dig deeper in investigating the basis for statistical claims in psychological research. For instance, if researchers are claiming that a particular measure is the same between different groups, are they basing this interpretation (incorrectly) on a statistically nonsignificant *p* value, or (correctly) on a Bayes factor? If the paper reports *p* values, what was the statistical power of the study? Ask the authors to estimate the chances that any obtained positive effects in their paper are *true* positives (i.e., the positive predictive value).
3. For journalists: As well as reporting the findings of published research, take into account its degree of transparency. Did the authors make their data publicly available? Did they preregister their study predictions before data collection? Did they publish their work through an OA pathway? Journalism has an important role to play in ensuring that scientists are held accountable for conducting transparent and reproducible research.

Coda

At the 2014 Association for Psychological Science conference in San Francisco, I gave a seminar on Registered Reports as part of a symposium called "The Replication Revolution: One Year On." It was a talk I had given many times before, but this time felt different. After over a year of giving presen-

tations on preregistration—mostly as an isolated voice—I now found that I was just one of many speakers advocating reform, among a lineup that included Bobbie Spellman, Brian Nosek, E. J. Wagenmakers, and Jelte Wicherts. There was also something quite unique about this meeting compared with the typical humdrum of an academic conference. Just hours earlier, some senior American psychologists had publicly accused proponents of open science of being "second stringers," "bullies," and "replication mafia." As the audience filed in, I wondered if that was really how psychologists saw us? Had the debate surrounding reform in psychology really become so politicized?

I have no idea how many psychologists attended that symposium—I would guess close to a thousand—but the atmosphere was electric. Spellman began by suggesting that psychology is in the midst of a political revolution, pushing transparency and reproducibility up the agenda to sit alongside the traditional aspirations of novelty and creativity. Along the way, attacks were to be expected; not everyone would agree with the current course—and in the questions that followed, it was clear that many did not. One year later, a physicist from CERN approached me after I gave an updated talk in London at a meeting of the Academy of Medical Sciences. Physics took this journey long ago, he said, and he was glad to see psychology moving the same way. "I am convinced that you are doing the right thing to turn your field into a good area of science," he later wrote. "I realise that you will have to fight off some nasty dinosaurs but I am convinced you will win—just look at what happened to the dinosaurs."

He was of course referring to the opponents of reform in psychology, Machiavelli's winners. But the truth is that the dinosaurs are within all of us—they are our unconscious biases, fragile egos, and propensity to cut corners. A dinosaur rears its head every time we are tempted to *p*-hack or change a hypothesis; every time we covet data as personal property; every time we judge each other based on superficial metrics. The dinosaur is the subjugation of our true mission to the ideals of storytelling, glory hunting, game playing and bean counting. Therefore, this book should not be seen as an attack on defenders of the status quo but as an intervention on ourselves, and a mission plan for self-improvement. As the saying goes, there is nothing noble in being superior to your fellow person—true nobility is being superior to your former self.

If we succeed then the future of psychological science is bright. We can look forward to deeper integration with computer science, biology, and physics, led by a culture of transparency that makes our work dependable and our data and tools reusable. Rather than looking back on psychology as a flawed indulgence, future generations of scientists will see the current period as a renaissance—a time when psychology left the cargo cult behind and became a truly rigorous and open science.

NOTES

Preface

1. For examples of psychology-related REF impact case studies, see "Assessing the viability of electric vehicles for daily use," http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=17132; "Facilitating intervention based on an enhanced understanding the antecedents and outcomes of debilitating exam-related anxiety," http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=43758; "Fundamental research on memory enables a robust criminal justice system," http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=30209; "Vision science and road Safety," http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=30208; "Human factors and space exploration," http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=27640; "Cardiff research supports the creation of the UK Climate Change Committee," http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=3480; "Influencing international tobacco policy on standardised tobacco packaging," http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=40192.

Chapter 1. The Sin of Bias

1. Daryl J. Bem, "Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect," *Journal of Personality and Social Psychology* 100, no. 3 (2011): 407, http://dx.doi.org/10.1037/a0021524.
2. Peter Aldhous, "Is this evidence that we can see the future?," *New Scientist*, 11 November 2010, www.newscientist.com/article/dn19712-is-this-evidence-that-we-can-see-the-future.html.
3. http://www.newscientist.com/article/dn20447-journal-rejects-studies-contradicting-precognition.html. French and Ritchie's nonreplication of Bem was eventually published in *PLOS ONE*: Stuart J. Ritchie, Richard Wiseman, and Christopher C. French, "Failing the future: Three unsuccessful attempts to replicate Bem's 'Retroactive Facilitation of Recall' Effect," *PLOS*

ONE 7, no. 3 (2012): e33423, http://dx.doi.org/10.1371/journal.pone.0033423.

4. Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, and Han L. J. van der Maas, "Why psychologists must change the way they analyze their data: The case of psi; Comment on Bem (2011)," (2011): 426, http://dx.doi.org/10.1037/a0022790. This article is available freely at http://www.ejwagenmakers.com/2011/WagenmakersEtAl2011_JPSP.pdf.
5. Georg Henrik von Wright, *A Treatise on Induction and Probability* (Oxfordshire: Psychology Press, 1951), 7: 86.
6. Peter C. Wason, "On the failure to eliminate hypotheses in a conceptual task," *Quarterly Journal of Experimental Psychology* 12, no. 3 (1960): 129–40, http://dx.doi.org/10.1080/17470216008416717.
7. Peter C. Wason, "Reasoning about a rule," *Quarterly Journal of Experimental Psychology* 20, no. 3 (1968): 273–81, http://dx.doi.org/10.1080/14640746808400161.
8. For a definitive review of confirmation bias and its suspected origins, see Raymond S. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises," *Review of General Psychology* 2, no. 2 (1998): 175.
9. Saul M. Kassin, Itiel E. Dror, and Jeff Kukucka, "The forensic confirmation bias: Problems, perspectives, and proposed solutions," *Journal of Applied Research in Memory and Cognition* 2, no. 1 (2013): 42–52, http://dx.doi.org/10.1016/j.jarmac.2013.01.001.
10. Alan G. Gross, "The roles of rhetoric in the public understanding of science," *Public Understanding of Science* 3, no. 1 (1994): 3–23, http://dx.doi.org/10.1088/0963–6625/3/1/001.
11. Joshua Klayman and Young-Won Ha, "Confirmation, disconfirmation, and information in hypothesis testing," *Psychological Review* 94, no. 2 (1987): 211, http://dx.doi.org/10.1037/0033–295X.94.2.211.
12. Hugo Mercier and Dan Sperber, "Why do humans reason? Arguments for an argumentative theory," *Behavioral and Brain Sciences* 34, no. 02 (2011): 57–74, http://dx.doi.org/10.1017/S0140525X10000968.
13. For early discussions of publication bias, the reader is directed to Theodore D. Sterling, "Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa," *Journal of the American Statistical Association* 54, no. 285 (1959): 30–34, http://dx.doi.org/10.1080/01621459.1959.10501497. The issue was revisited throughout the 1960s and 1970s, for instance by Robert Rosenthal in 1979: Robert Rosenthal, "The file drawer problem and tolerance for null results," *Psychological Bulletin* 86, no. 3 (1979): 638, http://dx.doi.org/10.1037/0033–2909.86.3.638.

In 1995, Sterling himself reassessed the problem and concluded that the "practice leading to publication bias have not changed over a period of 30 years." See Theodor D. Sterling, W. L. Rosenbaum, and J. J. Weinkam, "Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa," *American Statistician* 49, no. 1 (1995): 108–12, http://dx.doi.org/10.1080/00031305.1995.10476125.

14. Daniele Fanelli, " Positive" results increase down the hierarchy of the sciences," *PLOS ONE* 5, no. 4 (2010): e10068, http://dx.doi.org/10.1371/journal.pone.0010068.
15. www.nature.com/nature/authors/gta/others.html.
16. https://www.elsevier.com/journals/cortex/0010-9452/guide-for-authors.
17. http://brain.oxfordjournals.org/for_authors/general.html.
18. http://www.oxfordjournals.org/our_journals/cercor/for_authors/general.html.
19. Interview with outgoing editor of *Psychological Science*, Robert Kail: "Reflections on five years as editor," www.psychologicalscience.org/index.php/publications/observer/2012/november-12/reflections-on-five-years-as-editor.html.
20. Parts of this section have been adapted from an article on my personal blog: http://neurochambers.blogspot.co.uk/2012/03/you-cant-replicate-concept.html.
21. John A. Bargh, Mark Chen, and Lara Burrows, "Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action," *Journal of Personality and Social Psychology* 71, no. 2 (1996): 230, http://dx.doi.org/10.1037/0022–3514.71.2.230.
22. "Elderly-related words prime slow walking (#15)," http://www.psychfiledrawer.org/replication.php?attempt=MTU%3D.
23. "A primer for how not to respond when someone fails to replicate your work with a discussion of why replication failures happen" by Dan Simons: http://plus.google.com/+DanielSimons/posts/VJH8wXxxc3f.
24. Stéphane Doyen, Olivier Klein, Cora-Lise Pichon, and Axel Cleeremans, "Behavioral priming: It's all in the mind, but whose mind?," *PLOS ONE* 7, no. 1 (2012): e29081, http://dx.doi.org/10.1371/journal.pone.0029081.
25. John Bargh, "Nothing in their heads." The original post by Bargh was removed by *Psychology Today* but can still be read here: http://web.archive.org/web/20120308225833/http://www.psychologytoday.com/blog/the-natural-unconscious/201203/nothing-in-their-heads.
26. For Ed Yong's excellent coverage of this debate, see "A failed replication draws a scathing personal attack from a psychology professor," http://blogs.discovermagazine.com/notrocketscience/2012/03/10/failed-replication-bargh-psychology-study-doyen/.

27. Jay G. Hull, Laurie B. Slone, Karen B. Meteyer, and Amanda R. Matthews, "The nonconsciousness of self-consciousness," *Journal of Personality and Social Psychology* 83, no. 2 (2002): 406, http://dx.doi.org/10.1037/0022–3514.83.2.406.
28. Joseph Cesario, Jason E. Plaks, and E. Tory Higgins, "Automatic social behavior as motivated preparation to interact," *Journal of Personality and Social Psychology* 90, no. 6 (2006): 893, http://dx.doi.org/10.1037/0022–3514.90.6.893.
29. "Priming effects replicate just fine, thanks" by John Bargh: http://www.psychologytoday.com/blog/the-natural-unconscious/201205/priming-effects-replicate-just-fine-thanks.
30. "How valid are our replication attempts" by Rolf Zwaan: http://rolfzwaan.blogspot.be/2013/06/how-valid-are-our-replication-attempts.html.
31. For additional critique of conceptual replication, see Harold Pashler and Christine R. Harris, "Is the replicability crisis overblown? Three arguments examined," *Perspectives on Psychological Science* 7, no. 6 (2012): 531–36, http://dx.doi.org/10.1177/1745691612463401.
32. Norbert L. Kerr, "HARKing: Hypothesizing after the results are known," *Personality and Social Psychology Review* 2, no. 3 (1998): 196–217, http://dx.doi.org/10.1207/s15327957pspr0203_4.

    Although Kerr was the first to coin the term "HARKING," the concept had been proposed. Psychologist Adriaan de Groot (1956) was one of the first to warn against the dangers of hindsight bias in in hypothesis testing. See A. D. de Groot, "The meaning of 'significance' for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas]," *Acta Psychologica* 148 (2014): 188–94, http://dx.doi.org/10.1016/j.actpsy.2014.02.001.
33. Kerr, "HARKing," http://dx.doi.org/10.1207/s15327957pspr0203_4.
34. Leslie K. John, George Loewenstein, and Drazen Prelec, "Measuring the prevalence of questionable research practices with incentives for truth telling," *Psychological Science* (2012): 0956797611430953, http://dx.doi.org/10.1177/0956797611430953.
35. D. J. Bem, "Writing the empirical journal article," in Mark P. Zanna and John M. Darley, eds., *The Compleat Academic: A Practical Guide for the Beginning Social Scientist* (New York: Random House, 1987). An online version of this article is available here: http://web.archive.org/web/20140701082131/http://www.writingcenter.uconn.edu/pdf/Writing_the_Empirical_Journal_Article

_BEM.pdf. Similar sentiments were echoed in a 2012 online discussion by Fritz Strack of the University of Würzburg: "Also, as I do still not believe in ESP, I am not interested in the researchers' beliefs and their a priori hypotheses. I want to know what they learned and do not need to trace their many errors along the way. I am aware that this may deviate from the basic tenets of inferential statistics." https://groups.google.com/d/msg/spsp-discuss/JSUyofm1XDI/-uGnzzJTtHAJ.

Chapter 2. The Sin of Hidden Flexibility

1. For an excellent overview of NHST and common pitfalls, see the web resource "Statistics done wrong" by Alex Reinhart: http://www.statisticsdonewrong.com/index.html. See also Steven Goodman, "A dirty dozen: Twelve *p*-value misconceptions," *Seminars in Hematology* 45, no. 3 (2008): 135–40, http://dx.doi.org/10.1053/j.seminhematol.2008.04.003.
2. Ronald Fisher, "The arrangement of field experiments," *Journal of the Ministry of Agriculture* 33 (1926): 503–13. Available online at http://digital.library.adelaide.edu.au/dspace/bitstream/2440/15191/1/48.pdf.
3. Valen E. Johnson, "Revised standards for statistical evidence," *Proceedings of the National Academy of Sciences USA* 110, no. 48 (2013): 19313–17, http://dx.doi.org/10.1073/pnas.1313476110.
4. Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn, "False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychological Science* (2011): 1359–66, http://dx.doi.org/10.1177/0956797611417632.
5. Dorothy Bishop, "Interpreting unexpected significant results," http://deevybee.blogspot.co.uk/2013/06/interpreting-unexpected-significant.html.
6. Will M. Gervais and Ara Norenzayan, "Analytic thinking promotes religious disbelief," *Science* 336, no. 6080 (2012): 493–96, http://dx.doi.org/10.1126/science.1215647. A statistical critique can be found here: "Evidence of publication bias in 'Analytical thinking promotes religious disbelief,'" https://web.archive.org/web/20130817140545/http://jt512.dyndns.org/blog/?p=130. And see also "Do religious people lack analytical skills?" by Neuroautomaton, https://web.archive.org/web/20130408053541/http://neuroautomaton.com/religious-analytical-skills/.
7. E. J. Masicampo and Daniel R. Lalande, "A peculiar prevalence of *p* values just below. 05," *Quarterly Journal of Experimental Psychology* 65, no. 11 (2012): 2271–79, http://dx.doi.org/10.1080/17470218.2012.711335.

8. Nathan C. Leggett, Nicole A. Thomas, Tobias Loetscher, and Michael E. R. Nicholls, "The life of *p*: 'Just significant' results are on the rise," *Quarterly Journal of Experimental Psychology* 66, no. 12 (2013): 2303–9, http://dx.doi.org/10.1080/17470218.2013.863371.
9. Daniël Lakens from the Eindhoven University of Technology has argued that the spike in *p* values below .05, where it is shown to exist, could be due to factors other than *p*-hacking—most notably publication bias. Lakens suggests that rather than seeing a spike in *p* values immediately below .05 we are actually seeing a dip in those immediately *above* .05. This could arise owing to journals selectively rejecting studies that contain marginally nonsignificant results (e.g., $p = .06$). For more details, see Daniël Lakens, "On the challenges of drawing conclusions from *p*-values just below 0.05," *PeerJ* 3 (2015): e1142, http://dx.doi.org/10.7717/peerj.1142.
10. Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons, "*P*-curve: A key to the file-drawer," *Journal of Experimental Psychology: General* 143, no. 2 (2014): 534, http://dx.doi.org/10.1037/a0033242. The full text is freely available here: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2256237. The authors also provide an online resource: http://p-curve.com/.
11. When discussed at a scientific conference in 2012, *p*-curve generated an intense and heated debate, where (among other things) it was feared it could be misused as a weapon against individual researchers. The argument is summarized nicely in a blogpost by Sanjay Srivastava from the University of Oregon: "Does your *p*-curve weigh as much as a duck?," http://hardsci.wordpress.com/2012/02/10/does-your-p-curve-weigh-as-much-as-a-duck/.
12. Neuroskeptic, "False positive neuroscience," http://blogs.discovermagazine.com/neuroskeptic/2012/06/30/false-positive-neuroscience/.
13. For the study by Josh Carp, see Joshua Carp, "The secret lives of experiments: Methods reporting in the fMRI literature," *Neuroimage* 63, no. 1 (2012): 289–300, http://dx.doi.org/10.1016/j.neuroimage.2012.07.004.

    For discussion of likely false positive rates in fMRI research, see Tor D. Wager, Martin A. Lindquist, Thomas E. Nichols, Hedy Kober, and Jared X. Van Snellenberg, "Evaluating the consistency and specificity of neuroimaging data using meta-analysis," *Neuroimage* 45, no. 1 (2009): S210–S221, http://dx.doi.org/10.1016/j.neuroimage.2008.10.061.

    For discussion of fMRI reliability, see Craig M. Bennett and Michael B. Miller, "How reliable are the results from functional magnetic resonance imaging?," *Annals of the New York Academy of Sciences* 1191, no. 1 (2010): 133–55, http://dx.doi.org/10.1111/j.1749–6632.2010.05446.x, together with

a blog summary by Vaughan Bell: "How reliable are fMRI results?," at http://mindhacks.com/2010/03/04/how-reliable-are-fmri-results/.

In 2016, Russ Poldrack and colleagues published an authoritative overview of the threat posed to fMRI by low reproducibility and transparency. See Russell Poldrack, Chris I. Baker, Joke Durnez, Krzysztof Gorgolewski, Paul M. Matthews, Marcus Munafò, Thomas Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni, "Scanning the Horizon: Future challenges for neuroimaging research," *bioRxiv* (2016): 059188, http://biorxiv.org/content/early/2016/08/01/059188.

14. Uri Simonsohn vs. Norbert Schwarz, https://web.archive.org/web/20160516032001/http://opim.wharton.upenn.edu/~uws/SPSP/post.pdf; and see https://groups.google.com/forum/#!searchin/spsp-discuss/simonsohn/spsp-discuss/izOh5lrQDgQ/zLud5bZvPPUJ for a heated debate.
15. See "Replication, period. (A guest post by David Funder)" by David Funder, http://hardsci.wordpress.com/2012/09/21/replication-period-a-guest-post-by-david-funder/, and "Friday Fun: One researcher's *p*-curve analysis" by Michael Kraus, http://psych-your-mind.blogspot.co.uk/2012/02/friday-fun-one-researchers-p-curve.html.
16. As Sanjay Srivastava from the University of Oregon notes, failure to replicate a previous finding needn't indicate that the nonreplication was itself statistically different from the original finding. Showing this requires a direct test, which was satisfied in both nonreplications of the original Bargh study. http://hardsci.wordpress.com/2012/03/12/some-reflections-on-the-bargh-doyen-elderly-walking-priming-brouhaha/.
17. "Priming effects replicate just fine, thanks" by John Bargh: http://www.psychologytoday.com/blog/the-natural-unconscious/201205/priming-effects-replicate-just-fine-thanks.
18. The "one-time" prevalence estimate refers to the percentage of psychologists who have engaged in a particular practice on at least one occasion. It doesn't estimate wider prevalence.
19. Andrew Gelman and Eric Loken, "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or '*p*-hacking' and the research hypothesis was posited ahead of time," *Department of Statistics, Columbia University* (2013), http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.
20. Mark Stokes, "Biased debugging," http://the-brain-box.blogspot.co.uk/2013/02/biased-debugging.html. See also an interesting example of how such an error was caught by cognitive neuroscientist Russ Poldrack: "Anat-

omy of a coding error," http://www.russpoldrack.org/2013/02/anatomy-of-coding-error.html.

21. I am grateful to Eoin Travers (http://www.eointravers.com/) for making me aware of this term from studies of reasoning. For more information, see https://books.google.co.uk/books?id=iFMhZ4dl1KcC&lpg=PR9&ots=ZM1C6gn8p2&dq=evans%20newstead%20byrne&lr&pg=PA247#v=onepage&q=selective%20scrutiny%20account&f=false.
22. "Are research psychologists more like detectives or lawyers?" by John A. Johnson, http://www.psychologytoday.com/blog/cui-bono/201307/are-research-psychologists-more-detectives-or-lawyers-0.
23. Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn, "A 21 word solution," available at SSRN 2160588 (2012), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2160588.
24. See Michael J. Strube, "SNOOP: A program for demonstrating the consequences of premature and repeated null hypothesis testing," *Behavior Research Methods* 38, no. 1 (2006): 24–27, http://dx.doi.org/10.3758/BF03192746. Software is available here: http://web.archive.org/web/20160422220546/http://www.artsci.wustl.edu/~socpsy/Snoop.7z.

    For a similar proposal, see, Daniël Lakens, "Performing high-powered studies efficiently with sequential analyses," *European Journal of Social Psychology* 44, no. 7 (2014): 701–10, http://dx.doi.org/10.1002/ejsp.2023. The article can be downloaded freely from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2333729.
25. For a lucid introduction to Bayesian hypothesis testing, see Zoltan Dienes, "Bayesian versus orthodox statistics: Which side are you on?," *Perspectives on Psychological Science* 6, no. 3 (2011): 274–90, http://dx.doi.org/10.1177/1745691611406920. This paper can be freely downloaded from: http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/Dienes%202011%20Bayes.pdf.
26. Eric-Jan Wagenmakers, Michael Lee, Tom Lodewyckx, and Geoffrey J. Iverson, "Bayesian versus frequentist inference," in *Bayesian Evaluation of Informative Hypotheses*, 81–207 (New York: Springer, 2008), http://link.springer.com/chapter/10.1007/978–0–387–09612–4_9#page-1.
27. A. W. Edwards, *Likelihood: Expanded Edition* (Baltimore: Johns Hopkins University Press, 1992), 30. For a foundational and technically detailed text on the likelihood principle, the reader is directed also to: James O. Berger, Robert L. Wolpert, M. J. Bayarri, M. H. DeGroot, Bruce M. Hill, David A. Lane, and Lucien LeCam, "The likelihood principle," *Lecture Notes—Monograph Series* 6 (1988), http://web.uvic.ca/~dgiles/blog/Berger_and_Wolpert.pdf.

28. Dave Nussbaum, "The Stapel continuum," http://web.archive.org/web/20140604081238/http://www.davenussbaum.com/the-stapel-continuum/.

Chapter 3. The Sin of Unreliability

1. Geoff Brumfiel, "Particles break light speed limit," *Nature News*, 23 September 2011, http://www.nature.com/news/2011/110922/full/news.2011.554.html.
2. Robert Evans, "'Faster than light' particles threaten Einstein," *Reuters*, 23 September 2011, http://web.archive.org/web/20150321055654/http://www.reuters.com/article/2011/09/23/us-science-light-idUSTRE78L4FH20110923.
3. Michael D. Lemonick, "Was Einstein wrong? A faster-than-light neutrino could be saying yes," *Time*, 23 September 2011, http://content.time.com/time/health/article/0,8599,2094665,00.html.
4. Parts of this section have been adapted from: Chris Chambers, "Physics envy: Do 'hard' sciences hold the solution to the replication crisis in psychology?," *Guardian*, 10 June 2014, http://www.theguardian.com/science/head-quarters/2014/jun/10/physics-envy-do-hard-sciences-hold-the-solution-to-the-replication-crisis-in-psychology.
5. For the full text of Feynman's lecture, see http://web.archive.org/web/20160412070659/http://www.lhup.edu/~DSIMANEK/cargocul.htm.
6. Matthew C. Makel, Jonathan A. Plucker, and Boyd Hegarty, "Replications in psychology research: How often do they really occur?," *Perspectives on Psychological Science* 7, no. 6 (2012): 537–42, http://dx.doi.org/10.1177/1745691612460688.
7. Ioannidis reaches this estimation by taking into account the prior probability of a studied effect in psychology being nonnull (i.e, H0 is false), low rates of replication as a measure for ensuring self-correction, and the estimated rate of false positive discoveries. See John P. A. Ioannidis, "Why science is not necessarily self-correcting," *Perspectives on Psychological Science* 7, no. 6 (2012): 645–54, http://dx.doi.org/10.1177/1745691612464056.
8. Christopher J. Ferguson and Moritz Heene, "A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null," *Perspectives on Psychological Science* 7, no. 6 (2012): 555–61, http://dx.doi.org/10.1177/1745691612459059.
9. The complete special issue of replication studies can be freely downloaded from http://econtent.hogrefe.com/toc/zsp/45/3.
10. Simone Schnall's blog post can be read at http://web.archive

.org/web/20140531143904/http://www.spspblog.org/simone-schnall-on-her-experience-with-a-registered-replication-project/.

11. Dan Gilbert's extraordinary comment can be read at http://web.archive.org/web/20140531143904/http://www.spspblog.org/simone-schnall-on-her-experience-with-a-registered-replication-project/#comment-17137.
12. See http://web.archive.org/web/20141214024408/http://www.wjh.harvard.edu/~dtg/Bullies.pdf.
13. The language adopted by Gilbert and Schnall is reminiscent of immunologist Jacques Benveniste's reaction, in 1988, to the failed attempts to replicate his experiments on so-called water memory as the basis of homeopathy. Benveniste raised no objections to the replication methods before seeing the results. However, upon discovering that the data provided no evidence to support his "water memory" hypothesis, he condemned the entire process, writing in *Nature* that "Salem witchhunts or McCarthy-like prosecutions will kill science. Science flourishes only in freedom. We must not let, at any price, fear, blackmail, anonymous accusation, libel and deceit nest in our labs." http://web.archive.org/web/20070225102815/http://br.geocities.com/criticandokardec/benveniste02.pdf.
14. For the full text of Kahneman's letter, see http://www.scribd.com/doc/225285909/Kahneman-Commentary. In contrast to Kahneman's concerns about the reputational impact of nonreplications, a 2015 study by Adam Fetterman and Kai Sassenberg found that scientists overestimate the negative reputational impact of nonreplications. Moreover, they found that "admitting wrongness about a nonreplicated finding is less harmful to one's reputation than not admitting." Adam K. Fetterman and Kai Sassenberg, "The reputational consequences of failed replications and wrongness admission among scientists," *PLOS ONE* 10, no. 12 (2015): e0143723, http://dx.doi.org/10.1371/journal.pone.0143723.
15. Andrew Wilson, "Psychology's real replication problem: Our Methods sections," http://psychsciencenotes.blogspot.co.uk/2014/05/psychologys-real-replication-problem.html.
16. Jon Butterworth. Personal interview via e-mail. 2 June 2014.
17. Helen Czerski, personal interview via e-mail, 1 June 2014.
18. Katherine Mack. Personal interview via e-mail, 3 June 2014.
19. See http://journals.plos.org/plosone/s/criteria-for-publication#loc-2.
20. Jason Mitchell, "On the emptiness of failed replications," http://web.archive.org/web/20150604192510/http://wjh.harvard.edu/~jmitchel/writing/failed_science.htm.
21. For a range of critical responses to Jason's Mitchell essay, see Chris Said, "Jason Mitchell's essay," http://web.archive.org/web/20140916000231/http://

filedrawer.wordpress.com/2014/07/07/jason-mitchells-essay/; Sean Mackinnon, "Response to Jason Mitchell's 'On the emptiness of failed replications,'" http://web.archive.org/web/20141013180548/http://osc.centerforopenscience.org/2014/07/09/response-to-jason-mitchell/; Neuroskeptic, "On 'On the emptiness of failed replications,'" http://blogs.discovermagazine.com/neuroskeptic/2014/07/07/emptiness-failed-replications/; Bob Lewis, "In defense of replication studies," http://web.archive.org/web/20150904102838/http://trustinbob.blogspot.co.uk/2014/07/in-defense-of-replication-studies.html; Dorothy Bishop, "Replication and reputation: Whose career matters?," http://deevybee.blogspot.co.uk/2014/08/replication-and-reputation-whose-career.html.

22. Wolfgang Stroebe and Fritz Strack, "The alleged crisis and the illusion of exact replication," *Perspectives on Psychological Science* 9, no. 1 (2014): 59–71, http://dx.doi.org/10.1177/1745691613514450.
23. Daniel J. Simons, "The value of direct replication," *Perspectives on Psychological Science* 9, no. 1 (2014): 76–80, http://dx.doi.org/10.1177/1745691613514755. See also Stefan Schmidt, "Shall we really do it again? The powerful concept of replication is neglected in the social sciences," *Review of General Psychology* 13, no. 2 (2009): 90, http://dx.doi.org/10.1037/a0015108.
24. Jacob Cohen, "The statistical power of abnormal-social psychological research: A review," *Journal of Abnormal and Social Psychology* 65, no. 3 (1962): 145, http://dx.doi.org/10.1037/h0045186.
25. Peter Sedlmeier and Gerd Gigerenzer, "Do studies of statistical power have an effect on the power of studies?," *Psychological Bulletin* 105, no. 2 (1989): 309, http://dx.doi.org/10.1037/0033–2909.105.2.309.
26. Scott Bezeau and Roger Graves, "Statistical power and effect sizes of clinical neuropsychology research," *Journal of Clinical and Experimental Neuropsychology* 23, no. 3 (2001): 399–406, http://dx.doi.org/10.1076/jcen.23.3.399.1181.
27. Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò, "Power failure: Why small sample size undermines the reliability of neuroscience," *Nature Reviews Neuroscience* 14, no. 5 (2013): 365–76, http://dx.doi.org/10.1038/nrn3475.
28. Klaus Fiedler, Florian Kutzner, and Joachim I. Krueger, "The long way from α-error control to validity proper problems with a short-sighted false-positive debate," *Perspectives on Psychological Science* 7, no. 6 (2012): 661–69, http://dx.doi.org/10.1177/1745691612462587.
29. In brief, this is because when $p \cong .05$, the left half of the normal distribution (a symmetrical bell curve) that characterizes the observed effect will fall

within 1.96 standard deviations of the null distribution. Upon exact replication with the same sample size, there is therefore a roughly 50 percent chance of returning a nonsignificant test statistic ($p > .05$). Increasing the power requires doubling or tripling the original sample size. For a full explanation, see figure 1 in Button et al., "Power failure."

30. Free software packages such as G* Power now make prospective power analysis easy. http://www.gpower.hhu.de/en.html.
31. Scott E. Maxwell, "The persistence of underpowered studies in psychological research: Causes, consequences, and remedies," *Psychological Methods* 9, no. 2 (2004): 147, http://dx.doi.org/10.1037/1082–989X.9.2.147. An additional source of complexity in power analyses is deciding what effect size is appropriate. Because of publication bias, published effect sizes overestimate true effects sizes; therefore researchers should always power their studies according to the lower bound estimate of the likely effect size. This computation can be challenging and complex. For discussion, see Scott E. Maxwell, Michael Y. Lau, and George S. Howard, "Is psychology suffering from a replication crisis? What does 'failure to replicate' really mean?," *American Psychologist* 70, no. 6 (2015): 487, http://dx.doi.org/10.1037/a0039400.
32. Rachel A. Smith, Timothy R. Levine, Kenneth A. Lachlan, and Thomas A. Fediuk, "The high cost of complexity in experimental design and data analysis: Type I and type II error rates in multiway ANOVA," *Human Communication Research* 28, no. 4 (2002): 515–30, http://dx.doi.org/10.1111/j.1468–2958.2002.tb00821.x.
33. Etienne P. LeBel, Denny Borsboom, Roger Giner-Sorolla, Fred Hasselman, Kurt R. Peters, Kate A. Ratliff, and Colin Tucker Smith, "Psychdisclosure.org grassroots support for reforming reporting standards in psychology," *Perspectives on Psychological Science* 8, no. 4 (2013): 424–32, http://dx.doi.org/10.1177/1745691613491437.
34. This raises the question of why nearly half of published studies included measures that were apparently theoretically irrelevant. Given the high prevalence in psychology of post hoc hypothesizing (a.k.a. HARKing), it is perhaps more reasonable to assume that these measures were at least vaguely important at the outset of many studies but that the researchers decided after inspecting the results that they were no longer relevant to (or in agreement with) the HARKed hypothesis. As we saw in chapters 1 and 2, this form of cherry-picking violates the hypothetico-deductive model of the scientific method.
35. Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn, "A 21 word solution," available at SSRN 2160588 (2012), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2160588.

36. For two recent surveys showing how psychologists frequently misunderstand *p* values, see: Laura Badenes-Ribera, Dolores Frías-Navarro, Héctor Monterde-i-Bort, and Marcos Pascual-Soler, "Interpretation of the *p* value: A national survey study in academic psychologists from Spain," *Psicothema* 27, no. 3 (2015): 290–95, http://dx.doi.org/10.7334/psicothema2014.283; Laura Badenes-Ribera, Dolores Frias-Navarro, Bryan Iotti, Amparo Bonilla Campos, and Claudio Longobardi, "Misconceptions of the *p*-value among Chilean and Italian academic psychologists," *Frontiers in Psychology* 7 (2016): 1247, http://dx.doi.org/10.3389/fpsyg.2016.01247.
37. For an excellent and thorough list of misconceptions about *p* values, the reader is directed to see also Steven Goodman, "A dirty dozen: Twelve *p*-value misconceptions," *Seminars in Hematology* 45, no. 3 (2008): 135–40, http://dx.doi.org/10.1053/j.seminhematol.2008.04.003.
38. Sander Nieuwenhuis, Birte U. Forstmann, and Eric-Jan Wagenmakers, "Erroneous analyses of interactions in neuroscience: A problem of significance," *Nature Neuroscience* 14, no. 9 (2011): 1105–7, http://dx.doi.org/10.1038/nn.2886; Andrew Gelman and Hal Stern, "The difference between 'significant' and 'not significant' is not itself statistically significant," *American Statistician* 60, no. 4 (2006): 328–31, http://dx.doi.org/10.1198/000313006X152649. This article can be downloaded freely from http://www.stat.columbia.edu/~gelman/research/published/signif4.pdf.
39. For more information on the conditions that should lead to retraction of articles, see the Retraction Guidelines issued by the Committee on Publication Ethics: http://publicationethics.org/files/retraction%20guidelines.pdf.
40. Retraction due to failure to replicate is standard practice in sciences such as physics. Adam Marcus, "Doing the right thing: Authors retract lubricant paper whose findings they can't reproduce," http://retractionwatch.com/2014/03/14/doing-the-right-thing-authors-retract-lubricant-paper-whose-findings-they-cant-reproduce/. Similar examples can be found in neuroscience and biology. Ed Yong, "Narcolepsy paper retracted," http://phenomena.nationalgeographic.com/2014/07/30/narcolepsy-paper-retracted/.
41. Minhua Zhang and Michael L. Grieneisen, "The impact of misconduct on the published medical and non-medical literature, and the news media," *Scientometrics* 96, no. 2 (2013): 573–87, http://dx.doi.org/10.1007/s11192-012-0920-5.
42. Michael L. Grieneisen and Minghua Zhang, "A comprehensive survey of retracted articles from the scholarly literature," *PLoS One* 7, no. 10 (2012): e44118, http://dx.doi.org/10.1371/journal.pone.0044118. The quoted figure of 27 percent can be calculated from the data presented in figure S1 of this article by comparing the overall rate of retractions calculated across the sum

of all psychology disciplines (32 retractions per 50,974 articles = 0.063 percent) with the overall rate of retractions calculated across the sum of the remaining 233 disciplines (5,804 retractions per 2,501,816 articles = 0.23 percent). The rate of retractions in psychology (0.063 percent) is therefore 27 percent of the rate of retractions of the remaining fields (0.23 percent).

43. Dan Simons, "Replication, retraction, and responsibility," http://blog.dansimons.com/2014/01/replication-retraction-and.html.
44. John A. Bargh and Idit Shalev, "The substitutability of physical and social warmth in daily life," *Emotion* 12, no. 1 (2012): 154–62, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3406601/.
45. For a blog summary of the study, see http://traitstate.wordpress.com/2014/01/24/warm-water-and-loneliness/. The peer-reviewed replication is M. B. Donnellan, R. E. Lucas, and J. Cesario, "On the association between loneliness and bathing habits: Nine replications of Bargh and Shalev (2012) Study 1," *Emotion* 15, no. 1 (2015): 109, http://dx.doi.org/10.1037/a0036079. The article can be downloaded freely from https://msu.edu/~cesario/publications/Donnellan%20Lucas%20Cesario%20IN%20PRESS%20EMOTION%20loneliness%20bathing.pdf.
46. Will Gervais, "More power!," http://willgervais.com/blog/2014/3/5/more-power.
47. Eric-Jan Wagenmakers and B. U. Forstmann, "Rewarding high-power replication research," *Cortex* 51, no. 10 (2014), http://dx.doi.org/10.1016/j.cortex.2013.09.010.
48. Sanjay Srivastava, "A Pottery Barn rule for scientific journals," http://hardsci.wordpress.com/2012/09/27/a-pottery-barn-rule-for-scientific-journals/. We are planning to implement Sanjay's proposal for the first time in 2017 at the journal Royal Society Open Science. For details see: http://neurochambers.blogspot.co.uk/2016/11/an-accountable-replication-policy-at.html
49. Kimmo Eriksson and Brent Simpson, "Editorial decisions may perpetuate belief in invalid research findings," *PLOS ONE* 8, no. 9 (2013): e73364, http://dx.doi.org/10.1371/journal.pone.0073364.
50. Keith R. Laws, "Negativland—a home for all findings in psychology," *BMC Psychology* 1, no. 1 (2013): 2, http://dx.doi.org/10.1186/2050–7283–1–2.
51. This example is based on a real case of a reported test for Alzheimer's disease, critiqued by David Colquhoun: http://www.dcscience.net/?p=6473. See also this animation at the *Daily Mirror*: http://web.archive.org/web/20160206134529/http://ampp3d.mirror.co.uk/2014/03/11/how-a-90-accurate-alzheimers-test-can-be-wrong-92-of-the-time/. And for a technical overview, see http://www.biomedcentral.com/1741–7015/9/20.
52. Psychologist Gerd Gigerenzer has shown that the base rate fallacy can be largely overcome if problems are presented in terms of natural frequencies

("1 in 100") rather than probabilities ("1 percent"). Using a similar example to the one described in this chapter, but doing so in terms of natural frequencies, he found that correct answers among medical staff and students was as high as 76 percent. Rephrasing the current problem in terms of natural frequencies, as proposed by Gigerenzer, it would read like this:

> 10 out of 1000 people has Alzheimer's disease. A test has been developed to detect when a person has Alzheimer's disease. For 8 of the 10 people who have the disease, the test will come out positive. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for Alzheimer's disease. Imagine that we have assembled a random sample of 1000 people. They were selected by a lottery. Those who conducted the lottery had no information about the health status of any of these people. How many people who test positive for the disease will actually have the disease? ___ out of ___.

See Gerd Gigerenzer, "How to make cognitive illusions disappear: Beyond 'heuristics and biases,'" *European Review of Social Psychology* 2, no. 1 (1991): 83–115, http://dx.doi.org/10.1080/14792779143000033. The article can be downloaded freely at http://library.mpib-berlin.mpg.de/ft/gg/GG_How_1991.pdf.

53. In limited situations Bayes factors can even be multiplied together—for instance, suppose in two parallel experiments, values of $B = 1.6$ and $B = 2.5$ are uncovered in favor of H1 over H0. While neither value is sufficient to generate substantial evidence (>3) for H1, collectively they lead to $B = 4$ (1.6×2.5). It should be noted that the multiplication of Bayes factors is possible only when H1 is represented as a point hypothesis (i.e., a specific value) that is unchanged by the addition of new data. This might also apply in cases where two parallel (but independent) studies are run with identical methods and thus identical priors. In most cases, however, where a replication is conducted after an initial study, additional data will change the precise parameters of H1; in this case, the second analysis would adjust H1 based on all available data, preventing simple multiplication of the priors.
54. For an accessible introduction to Bayesian hypothesis testing in psychology, including worked-through examples, the reader is directed to two articles by Zoltan Dienes from the University of Sussex: Zoltan Dienes, "Bayesian versus orthodox statistics: Which side are you on?," *Perspectives on Psychological Science* 6, no. 3 (2011): 274–90, http://dx.doi.org/10.1177/1745691611406920, which can be freely downloaded from http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/Dienes%202011%20Bayes.pdf; and Zoltan Dienes, "Using Bayes to get the most out of non-significant results," *Frontiers in Psychology* 5,

no. 781 (2014): 1–17, http://dx.doi.org/10.3389/fpsyg.2014.00781. Other excellent resources include Alexander Etz's series of free and highly accessible blogposts, "Understanding Bayes": http://alexanderetz.com/understanding-bayes/; and Eric-Jan Wagenmakers, Richard D. Morey, and Michael D. Lee, "Bayesian benefits for the pragmatic researcher," https://osf.io/3tdh9/.

55. For an excellent example of adversarial collaboration, the reader is directed to a recent study of eye movements and memory recall: Dora Matzke, Sander Nieuwenhuis, Hedderik van Rijn, Heleen A. Slagter, Maurits W. van der Molen, and Eric-Jan Wagenmakers, "The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration," *Journal of Experimental Psychology: General* 144, no. 1 (2015): e1, http://dx.doi.org/10.1037/xge0000038.
56. Nature Neuroscience Editorial, "Raising standards," http://www.nature.com/neuro/journal/v16/n5/full/nn.3391.html.
57. For details on the initiative at *Psychological Science*, see https://www.psychologicalscience.org/index.php/publications/journals/psychological_science/ps-submissions#. See also http://www.theguardian.com/science/head-quarters/2014/jul/09/case-report-forms-psychology-replication for a call by psychologist Pete Etchells for psychologists to adopt more detailed approach to reporting methodology called "case report forms."
58. For a summary of the Center for Open Science badges initiative, see https://osf.io/tvyxz/wiki/home/.
59. Adam Marcus and Ivan Oransky, "Time for a reproducibility index," http://www.labtimes.org/labtimes/ranking/dont/2013_04.lasso.

Chapter 4. The Sin of Data Hoarding

1. For the NIH data-sharing policy, see http://www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html.
2. For *Science* magazine's requirements on data availability, see http://www.sciencemag.org/authors/science-editorial-policies#data-deposition.
3. For Earl Miller's tweet, see https://twitter.com/MillerLabMIT/status/360368532774592512.
4. Uri Simonsohn, "Just post it: The lesson from two cases of fabricated data detected by statistics alone," *Psychological Science* 24, no. 10 (2013): 1875–88, http://www.dx.doi.org/10.1177/0956797613480366.
5. http://www.nature.com/news/scientists-losing-data-at-a-rapid-rate-1.14416#/b1.
6. Jelte M. Wicherts and Marjan Bakker, "Publish (your data) or (let the data)

perish! Why not publish your data too?," *Intelligence* 40, no. 2 (2012): 73–76, http://dx.doi.org/10.1016/j.intell.2012.01.004.

7. A fine example is the work of Mark Stokes at the University of Oxford, who in 2013 reanalyzed a large archive of animal neurophysiology data to make a key discovery about working memory: Mark G. Stokes, Makoto Kusunoki, Natasha Sigala, Hamed Nili, David Gaffan, and John Duncan, "Dynamic coding for cognitive control in prefrontal cortex," *Neuron* 78, no. 2 (2013): 364–75, http://dx.doi.org/10.1016/j.neuron.2013.01.039.

   Another nice example is a 2012 study by Gilles Dutilh and colleagues. They took an existing data set that had been created to validate a new database of Dutch word frequencies and used it to test theories of why people are slower to make responses after committing errors in a decision-making task: Gilles Dutilh, Joachim Vandekerckhove, Birte U. Forstmann, Emmanuel Keuleers, Marc Brysbaert, and Eric-Jan Wagenmakers, "Testing theories of post-error slowing," *Attention, Perception, and Psychophysics* 74, no. 2 (2012): 454–65, http://dx.doi.org/10.3758/s13414–011–0243–2.
8. Heather A. Piwowar and Todd J. Vision, "Data reuse and the open data citation advantage," *PeerJ* 1 (2013): e175, http://dx.doi.org/10.7717/peerj.175.
9. Heather A. Piwowar, Roger S. Day, and Douglas B. Fridsma, "Sharing detailed research data is associated with increased citation rate," *PLOS ONE* 2, no. 3 (2007): e308, http://dx.doi.org/10.1371/journal.pone.0000308.
10. Jelte M. Wicherts, Denny Borsboom, Judith Kats, and Dylan Molenaar, "The poor availability of psychological research data for reanalysis," *American Psychologist* 61, no. 7 (2006): 726, http://dx.doi.org/10.1037/0003–066X.61.7.726. The article can be freely downloaded from https://web.archive.org/web/20150420101309/http://wicherts.socsci.uva.nl/datasharing.pdf
11. http://memforms.apa.org/apa/cli/interest/ethics1.cfm#8_14.
12. Jelte Wicherts, personal interview via e-mail, 28 August 2014.
13. Jelte M. Wicherts, Marjan Bakker, and Dylan Molenaar, "Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results," *PLOS ONE* 6, no. 11 (2011): e26828, http://dx.doi.org/10.1371/journal.pone.0026828.
14. Jelte Wicherts, personal interview via e-mail, 28 August 2014.
15. John A. Bargh and Idit Shalev, "The substitutability of physical and social warmth in daily life," *Emotion* 12, no. 1 (2012): 154–62, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3406601/.
16. Dan Simons, "Replication, retraction, and responsibility," http://blog.dansimons.com/2014/01/replication-retraction-and.html.
17. American Psychological Association, 2010, Ethical principles of psychologists and code of conduct, section 8.14, http://www.apa.org/ethics/code/principles.pdf.

18. Brent Donnellan, "What's the first rule about John Bargh's data?," http://traitstate.wordpress.com/2012/09/20/whats-the-first-rule-about-john-barghs-data/.
19. Dan Simons, "The fog of data—secrecy and science," http://blog.dansimons.com/2012/09/the-fog-of-data-secrecy-and-science.html.
20. Ed Yong, "The data detective," *Nature News*, 2 July 2012, http://www.nature.com/news/the-data-detective-1.10937; and see also Ed Yong, "Uncertainty shrouds psychologist's resignation," *Nature News*, 12 July 2012, http://www.nature.com/news/uncertainty-shrouds-psychologist-s-resignation-1.10968.
21. J.B.S. Haldane, "The faking of genetical results," *Eureka* 6 (1941), http://www.archim.org.uk/eureka/27/faking.html.
22. For the PLOS data availability policy, see http://journals.plos.org/plosone/s/data-availability.
23. DrugMonkey, "PLOS is letting the inmates run the asylum and this will kill them," https://drugmonkey.wordpress.com/2014/02/25/plos-is-letting-the-inmates-run-the-asylum-and-this-will-kill-them/.
24. rxnm, "PLOS's open data fever dream," https://rxnm.wordpress.com/2014/02/25/fan-fiction/
25. Erin C. McKiernan, "My concerns about PLOS's new open data policy," https://emckiernan.wordpress.com/2014/02/26/my-concerns-about-ploss-new-open-data-policy/.
26. Matthew D. MacManes, "Corner cases," http://genomebio.org/corner-cases/.
27. Some researchers question whether even standard anonymization is sufficient, as multiple anonymized data sets could, collectively, contain enough information to unblind participants.
28. At the time of writing, Figshare imposes a limit of 5GB per file, while Zenodo and Dataverse allow up to 2GB (though may permit larger file sizes on a case-by-case basis). Figshare offers individual user accounts for free but charges for institutional accounts.
29. See https://dx.doi.org/10.6084/m9.figshare.3381565.v1 for the analysis of the data.
30. Damian Pattinson, e-mail, 1 December 2015.
31. For guidelines at the journal *Cognition*, see http://www.elsevier.com/journals/cognition/0010–0277/guide-for-authors. For *Experimental Psychology*, see https://us.hogrefe.com/products/journals/exppsy. For *Archives of Scientific Psychology*, see http://www.apa.org/pubs/journals/arc/.
32. Researchers are divided on the issue of whether providing data, alone, is a significant enough contribution to justify coauthoship. In May 2016 I posed this question to Twitter, https://twitter.com/chrisdc77/status/7287096084

71179268. Overall, most respondents felt that sharing data did not justify authorship. Curiously, however, the community was more amenable to coauthorship in exchange for data when the data was shared privately between peers rather than publicly archived. Not sharing data was thus regarded as the more rewarded practice, suggesting that the lack of data archiving in psychology could be a form of rent seeking ("I'll give you access to my data if you make me an author on your paper").

33. For details on the badges initiative, see http://www.psychologicalscience.org/index.php/publications/journals/psychological_science/badges and https://osf.io/tvyxz/wiki/home/.
34. For further information, see http://centerforopenscience.org/top/.
35. See https://centerforopenscience.org/top/#list.
36. See http://www.esrc.ac.uk/files/about-us/policies-and-standards/esrc-research-data-policy/. One shortcoming of this policy is that data archiving is required only at the end of the project; therefore in most cases there is no link to the data in published peer-reviewed articles.
37. For more information, see https://opennessinitiative.org/.
38. Frederick Verbruggen, personal interview via e-mail, 22 September 2014.
39. D. V. M. Bishop, "Open research practices: Unintended consequences and suggestions for averting them (commentary on the peer reviewers' Openness Initiative)," *Royal Society Open Science* 3, no. 4 (2016): 160109, http://dx.doi.org/10.1098/rsos.160109. See also a provocative comment by Bishop and Stephan Lewandowsky, http://www.nature.com/news/research-integrity-don-t-let-transparency-damage-science-1.19219.
40. https://twitter.com/ImAlsoGreg/status/438179256284479488.
41. For a full outline of Candice and Richard's proposed "data partner scheme," see their blogpost "Habits and open data: Helping students develop a theory of scientific mind," http://bayesfactor.blogspot.no/2015/11/habits-and-open-data-helping-students.html.
42. Stephen Curry, personal interview via e-mail, 3 June 2014.

Chapter 5. The Sin of Corruptibility

1. In 2012, Stapel published a fascinating confessional of his rise and fall, entitled *Ontsporing* (which translates from Dutch as "Derailed" or "Derailment"). In 2014, psychologist Nick Brown translated the entire book into English with the new title *Faking Science: A True Story of Academic Fraud*. It can freely downloaded from http://web.archive

.org/web/20150325224546/https://errorstatistics.files.wordpress.com/2014/12/fakingscience-20141214.pdf.

2. There is a question mark over the exact point in time that Stapel began committing fraud. The Levelt Committee, which coordinated the formal investigation of his fraud, examined three periods in his career: University of Amsterdam, 1993–99 (including his PhD studies), the University of Groningen, 2000–2006 (where he became full professor); and Tilburg University, 2006–11 (where he established a research center for behavioral economics and became dean of research). Stapel claims that he used questionable practices from early on, but committed outright fabrication only from the Groningen period onward; however, the Drenth subcommittee, which was charged with investigating his earlier time in Amsterdam, found evidence of fraudulent practices as early as 1996. The veracity of Stapel's confession should therefore be viewed with a grain of salt. What seems clear is that Stapel escalated the frequency and severity of his data manipulation and fabrication as he progressed.
3. Stapel, *Faking Science*, trans. Brown, 100–101.
4. Ibid., 107.
5. Christopher J. Ferguson and Moritz Heene, "A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null," *Perspectives on Psychological Science* 7, no. 6 (2012): 555–61, http://dx.doi.org/10.1177/1745691612459059.
6. Stapel, *Faking Science*, trans. Brown, 109.
7. Ibid., 103.
8. Ibid., 101.
9. Ibid., 102.
10. Ibid., 103.
11. Ibid., 130.
12. Ibid., 128. There are two possible explanations for these illusory replications. Stapel may well have guessed correctly on some occasions because his hypotheses were often generated from a careful reading of the literature and so may have been true. Less optimistically, as we saw in chapter 1, psychology is plagued by confirmation bias in which researchers too easily *see what they want to see* in the data. Intense pressure is applied to produce results that agree with prior published effects, regardless of whether those effects are real.
13. The complete and final report of his fraud, published by the University of Tilburg, can be read here: https://www.tilburguniversity.edu/upload/3ff904d7–547b-40ae-85fe-bea38e05a34a_Final%20report%20Flawed%20Science.pdf.

14. For the interim report of the Stapel case, see http://web.archive.org/web/20160627142859/https://www.tilburguniversity.edu/upload/547aa461-6cd1-48cd-801b-61c434a73f79_interim-report.pdf.
15. Kate Kelland, "Dutch psychologist admits he made up research data," *Reuters*, 2 November 2011, http://www.reuters.com/article/us-dutch-scientist-fraud-idUSTRE7A12PL20111102.
16. Number 1 in this unenviable ranking is anesthesiologist Yoshitaka Fujii, who since 2012 has had 183 papers retracted for data fabrication. For the full list, see http://retractionwatch.com/the-retraction-watch-leaderboard/.
17. http://retractionwatch.com/2014/10/03/curtain-up-on-second-act-for-dutch-fraudster-stapel-college-teacher/#comment-31309.
18. http://retractionwatch.com/2014/10/03/curtain-up-on-second-act-for-dutch-fraudster-stapel-college-teacher/#comment-31388.
19. The report rather damningly notes, "Suspicions about data provided by Mr Stapel had also arisen among fellow full professors on two occasions in the past year. These suspicions were not followed up. The Committee concludes that the three young whistle-blowers showed more courage, vigilance and inquisitiveness than incumbent full professors" (46), https://www.tilburguniversity.edu/upload/3ff904d7–547b-40ae-85fe-bea38e05a34a_Final%20report%20Flawed%20Science.pdf.
20. Stapel, *Faking Science*, trans. Brown, 88.
21. Daniele Fanelli, "How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data," *PLOS ONE* 4, no. 5 (2009): e5738, http://dx.doi.org/10.1371/journal.pone.0005738.
22. Leslie K. John, George Loewenstein, and Drazen Prelec, "Measuring the prevalence of questionable research practices with incentives for truth telling," *Psychological Science* (2012): 0956797611430953, http://dx.doi.org/10.1177/0956797611430953.
23. Parts of this section appeared in a blog post at SciLogs: Chris Chambers, "Tackling the F Word," http://web.archive.org/web/20160316092653/http://www.scilogs.com/sifting_the_evidence/tackling-the-f-word/.
24. For the initial report on the Smeesters case, see http://www.eur.nl/fileadmin/ASSETS/press/2012/Juli/report_Committee_for_inquiry_prof._Smeesters.publicversion.28_6_2012.pdf.
25. Rolf Zwaan, personal interview via e-mail, 28 August 2015.
26. The violation of the law of small numbers in Smeesters's data uncovered by Uri Simonsohn (see chapter 4) provides evidence that the data was generated artificially, although this evidence fell short of providing a sufficient standard to officially conclude fraud. Zwaan told me that the committee couldn't see any way either *p*-hacking or sloppy practices could produce such

a pattern. "We tried to think of innocent copy and paste errors in Excel could have been the source of the problem. We couldn't think of any way in which this could happen but were also unable to completely rule it out."

27. http://retractionwatch.com/2014/03/19/final-report-in-smeesters-case-serves-up-seven-retractions/.
28. "Hoogleraar Erasmus ontkent fraude," http://www.nu.nl/algemeen/2844322/hoogleraar-erasmus-ontkent-fraude.html. Separately, Zwaan told me that Smeesters was probably encouraged to resign by Erasmus University. "I suspect he was pushed but this probably happened in his home department and/or at higher levels."
29. See http://dx.doi.org/10.6084/m9.figshare.3381574.v1. In the same interview, Smeesters admits that he would do things differently next time. "I do know that if I were to start over, I would deal with issues that are more relevant to society and that contribute to the wellbeing of people. Then, it would be of lesser importance whether or not something is innovative theoretically or statistically significant. Journals almost exclusively publish studies with statistically significant results, which can indeed lead to massaging your data."
30. For the 2012 internal report written by the whistle-blower in the Förster case, see https://retractionwatch.files.wordpress.com/2014/04/report_foerster.pdf.
31. Bruce S. Weir, "The rarity of DNA profiles," *Annals of Applied Statistics* 1, no. 2 (2007): 358, http://dx.doi.org/10.1214%2F07-AOAS128.
32. The whistle-blower told me: "[B]ecause all the data were gone, the committee could not investigate whether this was actual fraud; hence there was no verdict of any violation of scientific misconduct. . . . The highest administrators of UVA [University of Amsterdam], one of the biggest universities in the Netherlands, thought that because he'd lost all the data, he's excused."
33. For the English translation of the LOWI report, see https://retractionwatch.files.wordpress.com/2014/05/translation_lowi.pdf. For reasons that remain unclear, LOWI chose to scrutinize just one of the three publications highlighted by the whistle-blower—the article by Förster and Denzler (2012). The whistle-blower himself is unclear why this is, although it may have been a limitation imposed by the university. "I have an email of the former chair of the LOWI indicating that the University of Amsterdam only sent to the LOWI the data from the last paper, which is the only one of the three papers that had a co-author," he said. Thus, although Förster claimed that the raw data had been destroyed, summary data files were eventually provided for at least this one paper. However, the whistle-blower questions the veracity of this data, suggesting that during a period of sick leave, Förster had ample time to read up on the Stapel fraud case and learn how to beat the system in

a way that the university was unable or unwilling to prevent. "The data files were created about half a year after I filed the complaint. And created files, not 'saved as' files. . . . It was during [Förster's] 8 month sick leave. He had 7 months full time to make up the data. The data sets are completely useless. . . . Don't you think that if he had 7 or 8 months full time to save his career, that he would have read the Stapel report? He did his homework. The problem is the University of Amsterdam didn't investigate this, they completely refused it. It would have financial, reputational, consequences." In response to the allegation that he fabricated the data files after the fact, Förster claims that the files did indeed contain the original data, but that changing variable names from German to English led to the formation of a new data file and thus a data file that postdated the study in question: http://retractionwatch.com/2014/06/02/forster-on-defense-again-this-time-weighing-in-on-timeline-controversy/. At the time of writing, however, he has yet to publicly release this data or the original time-stamped files in German, stating "I thought about [posting the data], but my current experience with 'the net' prevents me from doing this. I will share the data with scientists who want to have a look at it and who are willing to share their results with me. But I will not leave it to an anonymous crowd that can post whatever it wants, including incorrect conclusions and insults." http://retractionwatch.com/2014/05/12/i-never-manipulated-data-forster-defends-actions-in-open-letter/.

34. "Social psychologist Förster denies misconduct, calls charge 'terrible misjudgment,'" http://retractionwatch.com/2014/04/30/social-psychologist-forster-denies-misconduct-calls-charge-terrible-misjudgment/.
35. "'I never manipulated data': Förster defends actions in open letter," http://retractionwatch.com/2014/05/12/i-never-manipulated-data-forster-defends-actions-in-open-letter/. He elaborated that his lack of clear records prevented his own investigation of who might be responsible. "During the time of [the] investigation I tried to figure out who could have done something inappropriate. However, I had to accept that there is no chance to trace this back; after all, the studies were run more than 7 years ago and I am not even entirely sure when, and I worked with too many people. I also do not want to point to people just because they are for some reason more memorable than others."
36. http://www.socolab.de/main.php?id=66. For additional summaries of the Förster affair, see http://osc.centerforopenscience.org/2014/05/29/forster-case/; and the Data Colada investigation, http://datacolada.org/2014/05/08/21-fake-data-colada/.
37. https://twitter.com/DegenRolf/status/643671007090339840. In November

2015, Förster agreed to retract two additional articles: http://retractionwatch.com/2015/11/12/psychologist-jens-forster-settles-case-by-agreeing-to-2-retractions/.

38. In 2013, the *New York Times* published an interview with Stapel in which the headline described his fraud as "audacious"—a synonym for "brave" and "daring." The headline was later changed to remove the term. http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html.
39. The full retraction notice can be read here: http://cercor.oxfordjournals.org/content/23/8/2015.full.
40. "Fraud topples second neuroscience word processing paper," http://retractionwatch.com/2014/01/13/fraud-topples-second-neuroscience-word-processing-paper/.
41. This was also confirmed by the University of Leuven Commission of Scientific Integrity, which appended the original retraction notice at *Cerebral Cortex* with a statement making clear that "the analysis of the data represented in this paper was manipulated intentionally by the first author: Wouter Braet." See http://cercor.oxfordjournals.org/content/24/1/280.full.
42. Hans Op de Beeck, personal interviews via e-mail, August–December 2015.
43. http://www.standaard.be/cnt/dmf20130816_00694853. I thank Frederick Verbruggen for supplying an English translation. In writing this book I contacted Braet, but he declined the invitation to be interviewed. Op de Beeck feels that this was an understandable reaction because, in his view, Braet's case was covered overly negatively by the Belgian media: "Even the so-called high-quality newspapers behaved as tabloids."
44. https://web.archive.org/web/20150907031741/https://ori.hhs.gov/content/case-summary-savine-adam-c.
45. http://ccpweb.wustl.edu/ORIresponse.html#13.
46. Remarkably, Braver was made aware of the Office of Research Integrity's conclusion of academic fraud only after it was made public. Speaking at the time to a journalist from the *St. Louis Post-Dispatch*, he said, "You learned of the outcome of Adam's case before I did, which I am pretty upset about—given that I was the one to report him in the first place." http://www.stltoday.com/lifestyles/health-med-fit/washington-u-student-s-mentor-talks-about-discredited-research/article_e2275d60–1ead-5906–851a-59c7a4daf6e5.html. In 2015, Braver told me that ORI have since revised their policy to keep principal investigators better informed.
47. http://ccpweb.wustl.edu//ORIresponse.html.
48. Todd Braver, personal interview via e-mail, 24 August 2015.

49. The society's full ethical policy can be found at https://www.sfn.org/member-center/professional-conduct/sfn-ethics-policy.
50. When I asked the ethics section of the Society for Neuroscience to comment on the ethical justification for their policy of collective punishment, my enquiry was redirected to their Communications and Marketing division. Their senior director of marketing stated that they "do not comment on individual cases" and referred me to their online ethics policy, which—as I made clear in my enquiry—makes no effort to justify the society's policy of collective punishment.
51. http://www.stltoday.com/lifestyles/health-med-fit/washington-u-student-s-mentor-talks-about-discredited-research/article_e2275d60-1ead-5906-851a-59c7a4daf6e5.html.
52. To ensure Kate's status as a protected whistle-blower, it is impossible to corroborate all of the content here with independent sources.
53. To protect Kate's identity within a fairly small research field, I cannot describe in detail the specific area in which she worked.

Chapter 6. The Sin of Internment

1. http://www.budapestopenaccessinitiative.org/read.
2. https://en.wikipedia.org/wiki/Creative_Commons_license. This ideal model of OA is called *libre* OA, in which users are granted rights not only to read the article but to reuse it.
3. Even traditional subscription publishing isn't necessarily free for authors. Many subscription journals, such as the *Journal of Neurophysiology*, levy page charges against authors—even when readers have to pay again to view content—and many journals in psychology and neuroscience include an additional fee to publish figures in color as opposed to black and white. As the saying goes, "there is always another fee."
4. Like many new journals, *Archives of Scientific Psychology* offers a temporary waiver on article processing charges, in this case until December 2016. At the time of writing, *BMC Psychology* charges authors US$2,145 per article, which is relatively high compared to other OA journals (e.g., *PLOS ONE*: US$1,495) and much higher than others (e.g., *PeerJ*: US$1,095). Some journals offer standing fee waivers or reductions for authors in low-income countries or for authors who are unable to cover the charges.
5. An additional nuance here is that a number of publishers permit immediate green OA on the author's personal website but impose a 12-month embargo for placing the same document in a public open access repository. In theory,

these forms of publishing are equally public; however, as the publishers are no doubt aware, the public repository is more likely to be picked up by automated search engines such as Google Scholar or a simple Internet search—thus the embargo serves the publisher's goal of temporarily limiting the visibility of the work in order to stimulate the market for sales. For a detailed and comprehensive database of publisher's policies regarding green and gold OA, the reader is directed to the excellent SHERPA/ROMEO website, http://www.sherpa.ac.uk/romeo/.

6. For a full directory of all OA journals, the reader is pointed to https://doaj.org/. It is notable that of more than 200 journals in this directory that are linked to psychology or neuroscience (journal title search for "psych* OR neur*" not one is regarded by the field as a prestigious and prominent outlet.
7. Parts of this section have appeared in the following *Guardian* blog: Chris Chambers, "Those who publish research behind paywalls are victims not perpetrators," http://www.theguardian.com/science/blog/2013/jan/23/open-access-publish-paywalls-victims-perpetrators.
8. ArXiv was launched in 1991 and now receives more than 8,000 submissions per month, http://arxiv.org/. A similar initiative has since been launched in biology (http://biorxiv.org/), and most recently in psychology (https://osf.io/view/psyarxiv). The OA journal *PeerJ* also offers a preprint archiving service.
9. Björn Brembs, Katherine Button, and Marcus Munafò, "Deep impact: Unintended consequences of journal rank," *Frontiers in Human Neuroscience* 7 (2013): 291, http://dx.doi.org/10.3389/fnhum.2013.00291.
10. For an impassioned argument on the morality of barrier-based publishing, see this *Guardian* blogpost by Mike Taylor: http://www.theguardian.com/science/blog/2013/jan/17/open-access-publishing-science-paywall-immoral.
11. For a damning analysis of this policy as launched at Queen Mary University London, the reader is directed to David Colquhoun's post, http://www.dcscience.net/2012/06/29/is-queen-mary-university-of-london-trying-to-commit-scientific-suicide/. Further examples of academic bean counting will be covered in chapter 7.
12. http://www.nature.com/news/funders-punish-open-access-dodgers-1.15007. The US National Institutes of Health have adopted a similar policy since 2008, which has experienced similar compliance problems despite sanctions.
13. http://www.wellcome.ac.uk/Managing-a-grant/End-of-a-grant/wtx026513.htm.
14. http://blog.wellcome.ac.uk/2015/03/03/the-reckoning-an-analysis-of-wellcome-trust-open-access-spend-2013-14/.

15. http://blog.wellcome.ac.uk/2015/10/22/10-years-of-open-access-at-the-wellcome-trust-in-10-numbers/.
16. The Finch Report: *Accessibility, Sustainability, Excellence: How to Expand Access to Research Publications*, 5. Available for free download at http://web.archive.org/web/20160322075144/http://www.researchinfonet.org/wp-content/uploads/2012/06/Finch-Group-report-FINAL-VERSION.pdf.
17. Ibid., 102.
18. Ibid., 7. One problem with gold OA is the question of who pays when neither the institution nor the researcher have any funds available. This can be a particular problem for early-career researchers based at less lucrative institutions. As a colleague put it: "My institution has said that they won't pay, so I've been stuck with the *PLOS ONE* fee. I asked them to waive it, but all they could do is reduce it. That's still $400 I have to find from somewhere. I'm still an early career researcher, so I don't have any slush funds to pay for it."
19. Ibid., 6.
20. Ibid., 36.
21. Harnad's full comment can be read at https://web.archive.org/web/20160418152013/http://openaccess.eprints.org/index.php?/archives/904-Finch-Report,-a-Trojan-Horse,-Serves-Publishing-Industry-Interests-Instead-of-UK-Research-Interests.html.
22. Stevan Harnad has referred to this double-dipping model as "fool's gold OA": http://openaccess.eprints.org/index.php?/archives/1007-Pre-Green-OA-Fools-Gold-vs.-Post-Green-OA-Fair-Gold.html.
23. http://www.rcuk.ac.uk/research/openaccess/policy/.
24. "From 'as soon as possible' to 'immediate' open access," http://www.nwo.nl/en/news-and-events/news/2015/from-as-soon-as-possible-to-immediate-open-access.html
25. "Open access in the next Research Excellence Framework: policy adjustments and qualifications," http://www.hefce.ac.uk/media/HEFCE,2014/Content/Pubs/2015/CL,202015/Print-friendly%20version.pdf.
26. Since 2012 the European Research Council has implemented an open access requirement on all publications arising from its grants, including psychology. It's policy requires that researchers:

    > (a) as soon as possible and at the latest on publication, deposit a machine-readable electronic copy of the published version or final peer-reviewed manuscript accepted for publication in a repository for scientific publications. Moreover, the beneficiary must aim to deposit at the same time the research data needed to validate the results presented in the deposited scientific publications.

(b) ensure open access to the deposited publication—via the repository—at the latest:

(i) on publication, if an electronic version is available for free via the publisher, or

(ii) within six months of publication (twelve months for publications in the social sciences and humanities) in any other case.

(c) ensure open access—via the repository—to the bibliographic metadata that identify the deposited publication, which must include a persistent identifier.

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf.

27. Andrea Kuszewski's tweet from 2011 that started the #icanhazpdf phenomenon: https://twitter.com/AndreaKuszewski/status/28257118322688000.
28. For an insightful analysis of the #icanhazpdf phenomenon, see Carolyn Caffrey Gardner and Gabriel J. Gardner, "Bypassing interlibrary loan via twitter: An exploration of# icanhazpdf requests," *ACRL (Association of College and Research Libraries)* (2015), http://web.archive.org/web/20160702052610/http://eprints.rclis.org/24847/2/gardner.pdf. During a three-month period in 2014, the authors found 824 #icanhazpdf requests in the Twitter archives, which extrapolates to approximately 3,296 requests over 12 months. Most requests (73 percent) were in the social and life sciences (including psychology).
29. At the time of writing, Sci-Hub could be found under several domains, including http://sci-hub.bz/ and http://sci-hub.cc/.
30. In particular, Elbakyan argues that publishers such as Elsevier stand in violation of the following clause of Article 27: "(1) Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits." http://www.un.org/en/universal-declaration-human-rights/.
31. "Sci-Hub tears down academia's 'illegal' copyright paywalls," https://torrentfreak.com/sci-hub-tears-down-academias-illegal-copyright-paywalls-150627/.
32. In 2014, Elsevier—which is the largest publisher of academic journals—was reported to return a 37 percent profit margin, equating to over $US1 billion, http://web.archive.org/web/20160315064211/https://libraries.mit.edu/scholarly/mit-open-access/open-access-at-mit/mit-open-access-policy/publishers-and-the-mit-faculty-open-access-policy/elsevier-fact-sheet/.
33. The full complaint from Elsevier can be read here: https://torrentfreak.com/images/elsevier-complaint.pdf.

34. "Court orders shutdown of Libgen, Bookfi and Sci-Hub," https://torrentfreak.com/court-orders-shutdown-of-libgen-bookfi-and-sci-hub-151102/.
35. "Sci-Hub, BookFi and LibGen resurface after being shut down," https://torrentfreak.com/sci-hub-and-libgen-resurface-after-being-shut-down-151121/.
36. For details of the indictment against Swartz, see http://www.wired.com/images_blogs/threatlevel/2012/09/swartzsuperseding.pdf.
37. The poorest institutions and countries are somewhat protected from this problem. Several major publishers participate in the Research4Life program, which at the time of writing offers free journal access to 72 "Group A" countries in the developing world, http://www.research4life.org/eligibility/.
38. In one of many such examples, Ross Mounce, a postdoctoral researcher at the Natural History Museum in London, recounts his experience of discovering that the University of Bath cut its subscription to the popular Royal Society journal, *Biology Letters*, http://rossmounce.co.uk/2012/02/12/journal-mega-bundles-thecostofknowledge/. In this case it appears the subscription was axed not only to save money but also because it didn't belong to a purchased "bundle." The largest for-profit publishers, such as Elsevier, often sell journal subscriptions to universities in large bundles and do not allow the libraries to selectively unsubscribe from specific journals. This means that when the purse strings are tightened, more isolated subscriptions—such as those to learned societies—can be more vulnerable to cuts (although note that at the time of writing, the University of Bath has since restored its subscription to *Biology Letters*). The details of library subscriptions, and what they cost, are often treated as commercial secrets, although a 2014 study by Theodore Bergstrom and colleagues was able to glean information from US universities using Freedom of Information requests: Theodore C. Bergstrom, Paul N. Courant, R. Preston McAfee, and Michael A. Williams, "Evaluating big deal journal bundles," *Proceedings of the National Academy of Sciences USA* 111, no. 26 (2014): 9425–30, http://dx.doi.org/10.1073/pnas.1403006111. In 2015, Mounce conducted a poll on social media to discover how many leading universities could access a randomly selected article from 1949—an article that was, in theory, available electronically. Of 70 institutions in the UK and North America, just 10 reported that they could successfully download the article: http://rossmounce.co.uk/2015/09/16/who-actually-has-access-to-paywalled-research/.
39. David Halpern, "Presidential column: Applying psychology to public policy," http://www.psychologicalscience.org/index.php/publications/observer/2014/january-14/applying-psychology-to-public-policy.html.
40. http://www.behaviouralinsights.co.uk/.

41. "Executive order—using behavioral science insights to better serve the American people," https://www.whitehouse.gov/the-press-office/2015/09/15/executive-order-using-behavioral-science-insights-better-serve-american.
42. POST is an in-house group of civil servants that provides independent, in-depth scientific briefings called POSTNotes to members of Parliament, http://www.parliament.uk/post.
43. Adriana De Palma, "Why all PhD students should do a policy placement," http://web.archive.org/web/20160601081827/https://therostrumblog.wordpress.com/2015/01/12/why-all-phd-students-should-do-a-policy-placement/.
44. https://twitter.com/stevenhill/status/670645979167748100.
45. https://twitter.com/ersatzben/status/670939925106302977.
46. https://twitter.com/ersatzben/status/613782419011989504.
47. Ben Johnson, "Subscription publishers do not want my business," http://web.archive.org/web/20160409024908/https://ersatzben.com/2015/06/20/subscription-publishers-do-not-want-my-business/. A similar case was documented by former *BMJ* editor Richard Smith, http://web.archive.org/web/20160414122448/http://blogs.bmj.com/bmj/2012/06/28/richard-smith-a-bad-bad-week-for-access/.
48. Chris Hartgerink, "Why I content mine," http://web.archive.org/web/20160612071438/http://onsnetwork.org/chartgerink/2015/11/19/why-i-content-mine/.
49. Chris Hartgerink, "Elsevier stopped me doing my research," http://web.archive.org/web/20160611032746/http://onsnetwork.org/chartgerink/2015/11/16/elsevier-stopped-me-doing-my-research/. One of the most remarkable aspects to this case is that Hartgerink's activities appeared to be assumed, without evidence, to be "stealing" of content. In an e-mail forwarded to Hartgerink, a New York employee of Elsevier allegedly wrote: "I have noticed what seems to be crawling at Tilburg recently. Recently we are seeing a lot of content theft by people who steal legitimate IDs to gain access to university resources. Will you please reach out to your contacts at Tilburg to ask them to investigate." Subsequently, the Dutch account manager contacted Tilburg (in Dutch, translated by Hartgerink) and allegedly wrote: "Given below is serious amounts of downloading which could indicate potential abuse [of access]. Could I ask you to investigate as soon as possible what is going on? If the downloads continue at this scale the access will be blocked for security reasons."
50. http://web.archive.org/web/20160611032746/http://onsnetwork.org/chartgerink/2015/11/16/elsevier-stopped-me-doing-my-research/#comment-11.
51. Chris Shillum, "Elsevier updates text-mining policy to improve access for re-

searchers," http://www.elsevier.com/connect/elsevier-updates-text-mining-policy-to-improve-access-for-researchers.

52. Chris Hartgerink, "Why Elsevier's 'solution' is the problem," http://web.archive.org/web/20160611010128/http://onsnetwork.org/chartgerink/2015/11/20/why-elseviers-solution-is-the-problem/.
53. The full memo is no longer available from Harvard University but can be read at http://web.archive.org/web/20160317160330/http://gantercourses.net/wp-content/uploads/2013/11/Faculty-Advisory-Council-Memorandum-on-Journal-Pricing-%C2%A7-THE-HARVARD-LIBRARY.pdf.
54. https://osc.hul.harvard.edu/policies/.
55. "Dutch universities start their Elsevier boycott plan," https://universonline.nl/2015/07/02/dutch-universities-start-their-elsevier-boycott-plan.
56. Daniel Allington, "On open access, and why it's not the answer," http://web.archive.org/web/20160402004400/http://www.danielallington.net/2013/10/open-access-why-not-answer/.
57. http://en.wikipedia.org/wiki/Out-group_homogeneity.
58. For an excellent overview of the benefits of open access publishing to those outside academia, see http://whoneedsaccess.org/.
59. John Bargh, "Nothing in their heads," http://web.archive.org/web/20120309203900/http://psychologytoday.com/blog/the-natural-unconscious/201203/nothing-in-their-heads.
60. Stéphane Doyen, Olivier Klein, Cora-Lise Pichon, and Axel Cleeremans, "Behavioral priming: It's all in the mind, but whose mind?," *PLOS ONE* 7, no. 1 (2012): e29081, http://dx.doi.org/10.1371/journal.pone.0029081.
61. http://journals.plos.org/plosone/s/criteria-for-publication.
62. Bohannon's report can read freely at http://www.sciencemag.org/content/342/6154/60.full.
63. https://doaj.org/.
64. Michael Eisen, "I confess, I wrote the arsenic DNA paper to expose flaws in peer-review at subscription based journals," http://web.archive.org/web/20160508112019/http://www.michaeleisen.org/blog/?p=1439.
65. C. Hajjem, Y. Gingras, and S. Harnad, "Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact," *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 28 no. 4 (2005): 39–46, http://sites.computer.org/debull/A05dec/hajjem.pdf.
66. For details, see http://sparceurope.org/oaca/.
67. For an insightful and authoritative review of the benefits of open access publishing, the reader is directed to Jonathan P. Tennant, François Waldner,

Damien C. Jacques, Paola Masuzzo, Lauren B. Collister, and Chris H. J. Hartgerink, "The academic, economic and societal impacts of Open Access: An evidence-based review," *F1000Research* 5 (2016), http://dx.doi.org/10.12688%2Ff1000research.8460.2.

Chapter 7. The Sin of Bean Counting

1. Goodhart's law is named after the economist Charles Goodhart, although the formulation quoted here has been attributed to anthropologist Marilyn Strathern. http://www.atm.damtp.cam.ac.uk/mcintyre/papers/LHCE/goodhart.html.
2. Eugene Garfield, "Citation indexes for science," *Science* 122 (1955): 108–11, http://www.garfield.library.upenn.edu/papers/science_v122v3159p108y1955.html.
3. Editorial: *Scientometrics* 1, no. 1 (1978): 3–8, http://dx.doi.org/10.1007/BF02016836.
4. E. Garfield, "The history and meaning of the journal impact factor," *JAMA* 295 no. 1 (2006): 90–93. The article can be freely accessed here: http://garfield.library.upenn.edu/papers/jamajif2006.pdf.
5. Papers can differ widely in the number of citations they receive for reasons unrelated to study quality. Articles can be highly cited to signpost a notable flaw or because of the prominence of the journal in which the article appeared. Studies can also be popular because they are novel (even if low quality) or because of impact in the print and broadcast media. On top of this, different scientific disciplines, and subdisciplines, have different baselines for what constitute a typical number of citations, making it difficult to draw comparisons. For an insightful discussion of these problems and more, see Robert Adler, John Ewing, and Peter Taylor, "Joint committee on quantitative assessment of research: Citation statistics," *Australian Mathematical Society Gazette* 35, no. 3 (2008): 166–88, http://www.austms.org.au/Gazette/2008/Jul08/Gazette35%283%29Web.pdf#page=24.
6. See Per O. Seglen, "Why the impact factor of journals should not be used for evaluating research," *BMJ: British Medical Journal* 314, no. 7079 (1997): 498, which showed no correlation between JIF and citation rate, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2126010/pdf/9056804.pdf. See also George A. Lozano, Vincent Larivière, and Yves Gingras, "The weakening relationship between the impact factor and papers' citations in the digital age," *Journal of the American Society for Information Science and Technology* 63, no. 11 (2012): 2140–45, which found a low correlation that is diminishing over

time, http://dx.doi.org/10.1002/asi.22731. This article can be freely downloaded from http://www.ost.uqam.ca/Portals/0/docs/articles/2012/JASISTno11.pdf.

7. As a personal example, one of my most highly cited articles, at the time of writing, is published in the specialist *Journal of Cognitive Neuroscience* in 2006 (371 cites; JIF = 3.6). By comparison, a similar experimental paper that I published two years earlier in the "high impact" journal *Nature Neuroscience* has less than half the number of citations (176 cites; JIF = 16.7).
8. Ferric C. Fang, R. Grant Steen, and Arturo Casadevall, "Misconduct accounts for the majority of retracted scientific publications," *Proceedings of the National Academy of Sciences USA* 109, no. 42 (2012): 17028–33, http://dx.doi.org/10.1073/pnas.1212247109. See also a summary of this and other problems with JIF on Björn Bremb's blog: http://blogarchive.brembs.net/news.php?item.766.11.
9. See *PLOS Medicine* editors, "The impact factor game," *PLOS Med* 3, no. 6 (2006): e291, http://dx.doi.org/10.1371/journal.pmed.0030291.
10. See Björn Brembs, Katherine Button, and Marcus Munafò, "Deep impact: Unintended consequences of journal rank," *Frontiers in Human Neuroscience* 7 (2013): 291, http://dx.doi.org/10.3389/fnhum.2013.00291.
11. Eugene Garfield, "Interview with Eugene Garfield, chairman emeritus of the Institute for Scientific Information (ISI)," *Cortex* 37, no. 4 (2001): 575–77. The article can be freely accessed at http://www.garfield.library.upenn.edu/papers/cortex37%284%292001.pdf.
12. At some Dutch universities there is an unofficial rule that students are expected to publish a minimum number of articles in peer-reviewed journals (usually at least two) before they can pass their PhD. On the positive side, this ensures that the student gains experience in the practice of scientific publishing; however, the much larger negative is that setting a minimum *quantity* also encourages small, underpowered studies and the salami slicing of work into minimal publishable units. Because of publication bias, it also increases the pressure on PhD students to engage in biased research practices (such as *p*-hacking) to ensure that their experiments produce positive, publishable results.
13. http://www.ascb.org/dora/. The Association of Dutch Universities, of which the University of Maastricht is a member, is organizational signatory 561.
14. See Brembs, Button, and Munafò, "Deep impact," 291, http://dx.doi.org/10.3389/fnhum.2013.00291.
15. Stephen Curry, "Sick of impact factors," http://occamstypewriter.org/scurry/2012/08/13/sick-of-impact-factors/.
16. Henry L. Roediger III, "Journal impact factors: How much should we care?,"

http://www.psychologicalscience.org/index.php/publications/observer/2013/september-13/journal-impact-factors.html.

17. https://uk.sagepub.com/en-gb/eur/psychological-science-package/journal201991.
18. https://twitter.com/CT_Bergstrom/status/657221745628200960.
19. http://www.ucl.ac.uk/hr/docs/proms/SnrProm_AnnexA.doc. In the event of this link being taken down, an archived copy can be found at https://web.archive.org/web/20160516082534/http://www.ucl.ac.uk/hr/docs/proms/SnrProm_AnnexA.doc.
20. For a recent example of how the Times Higher Education reinforces the idea that grants are prizes, see https://www.timeshighereducation.com/news/grant-winners-3-december-2015.
21. Dorothy Bishop has delivered a blistering attacking on the absurdity of evaluating scientists based on their grant income: http://deevybee.blogspot.co.uk/2014/12/why-evaluating-scientists-by-grant.html.
22. In general there is very little assessment of whether grants awarded to psychologists are well spent. One exception is the UK Economic and Social Research Council's rapporteur program, which asks reviewers to assess the "scientific, economic and societal impact" of outputs stemming from awarded grants, http://www.esrc.ac.uk/files/funding/guidance-for-peer-reviewers/guidance-notes-on-the-rapporteur-quality-and-impact-comment-form/. However, it is unclear how the feedback is used, and whether the funder uses the information from the rapporteur reports to inform future funding decisions.
23. In psychology it is possible to conduct entire research programs on the basis of little-to-no grant funding, for instance by conducting research projects with undergraduate research trainees or PhD students.
24. I put this scenario to a Twitter poll in December 2015, "Jane & Mary compete for academic promotion. They're matched in every way but Mary has grants. Who should/would get promoted?" Of 98 respondents, 57 percent believed that "Mary should and would" be promoted over Mary, while only 6 percent felt the "Jane should and would." Interestingly, a substantial number (33 percent) believed that "Mary would but Jane should," perhaps reflecting some recognition that inputs (grants) and outputs (publications, influence) should be weighed against each other rather than summed. Very few respondents (4 percent) felt that "Jane would but Mary should." Results of Twitter polls should be viewed with caution—this was not a properly sampled survey and the background of the respondents is unknown, https://twitter.com/chrisdc77/status/677829444103442432.

25. UCL pharmacologist David Colquhoun has doggedly investigated and exposed these cases. http://www.dcscience.net/2012/06/29/is-queen-mary-university-of-london-trying-to-commit-scientific-suicide/.
26. David Colquhoun, "Bad financial management at Kings College London means VC Rick Trainor is firing 120 scientists," http://www.dcscience.net/2014/06/07/bad-financial-management-at-kings-college-london-means-vc-rick-trainor-is-firing-120-scientists/.
27. John Allen and Fanis Missirlis's critique of Queen Mary's retention policy can be read freely at the *Lancet*: John F. Allen and Fanis Missirlis, "Queen Mary: Nobody expects the Spanish Inquisition," *Lancet* 379, no. 9828 (2012): 1785, http://www.thelancet.com/journals/lancet/article/PIIS0140–6736%2812%2960697–7/fulltext.
28. The COPE guidelines can be freely read at http://publicationethics.org/files/International%20standards_authors_for%20website_11_Nov_2011_0.pdf.
29. Like all rules of thumb surrounding authorship, there are many exceptions. For instance, in cases where two senior researchers work on a project together, it is not uncommon for one researcher to take first authorship and the other to take last authorship, despite both authors being of equal seniority. Another caveat with last authorship is that its value is context dependent: for researchers who are demonstrably independent, or on the cusp of independence, last authorship is sign of academic leadership. But junior scholars are generally better off settling for a second authorship compared with a last authorship. If the community knows you are too junior to be a principal investigator, a last authorship can be interpreted literally as contributing the least to a study. The whims and fancies of such conventions deliver another nail in the coffin to authorship rank as a marker of scientific contribution.
30. For the American Psychological Association's guidelines on authorship, see http://www.apa.org/research/responsible/publication/.
31. For the ICMJE guidelines on authorship, see http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html.
32. My article, entitled "Tough love II: 25 tips for early-career scientists": http://neurochambers.blogspot.co.uk/2012/05/tough-love-ii-25-point-guide-for-early.html.
33. Dorothy Bishop's well-aimed critique of my post can be read here: http://neurochambers.blogspot.co.uk/2012/05/tough-love-ii-25-point-guide-for-early.html?showComment=1338490130354#c1743257197227672347.

34. My rather weak response to Dorothy: http://neurochambers.blogspot.co.uk/2012/05/tough-love-ii-25-point-guide-for-early.html?showComment=1338498718060#c5457911252285746290.
35. Uta Frith's manifesto for slow science is pure wisdom. See http://frithmind.org/socialminds/2015/10/11/slow-science/.
36. This isn't always the case. It is not uncommon for papers to be cited inappropriately based on superficial or erroneous reading of the work by the authors.
37. Science is awash with metrics in addition to JIF. The Eigenfactor seeks to quantify the citation impact of a journal by considering not just the number of citations the articles receive but also the impact of the journal in which the citing article occurs. The *h*-index (named after its creator, Jorge E. Hirsch) attempts to reduce both the productivity and impact of individual scientists into a single number. A scientist is said to have an *h*-index of 20 if he or she has 20 articles that are cited at least 20 times each. In addition to focus solely on citations, the *h*-index has been criticized for favoring senior over junior scientists, failing to distinguish individual contributions to authorship, and being vulnerable to manipulation through self-citation. So-called altmetrics are another burgeoning set of measures that seek to quantify the impact of academic work in the print media, blogosphere, and social media. All such metrics tell us something, but none are suitable for expert judgments of quality and long-term consideration of how published research impacts theory and practice.

Chapter 8. Redemption

1. It should also be pointed out that just because an area of research doesn't meet all these conditions for being "science" doesn't mean that it is flawed, useless, or not research. Qualitative psychological research, for instance, can reveal rich insights into behavior and society even if such approaches do not seek to test hypotheses or quantify phenomena.
2. For a discussion of wider reproducibility problems in biomedicine, see the 2015 report issued by the UK Academy of Medical Sciences: http://www.acmedsci.ac.uk/policy/policy-projects/reproducibility-and-reliability-of-biomedical-research/.
3. Parts of this section are adapted from the following articles: Chris Chambers, "Are we finally getting serious about fixing science?," http://www.theguardian.com/science/head-quarters/2015/oct/29/are-we-finally-getting-serious-about-fixing-science; and Christopher D. Chambers, Eva Feredoes,

Suresh Daniel Muthukumaraswamy, and Peter Etchells, "Instead of 'playing the game' it is time to change the rules: Registered reports at AIMS Neuroscience and beyond," *AIMS Neuroscience* 1, no. 1 (2014): 4–17. This article can freely downloaded from http://www.aimspress.com/article/10.3934/Neuroscience.2014.1.4/pdf.

4. Chris Chambers, "Why I will no longer review or publish for the journal *Neuropsychologia*," http://neurochambers.blogspot.co.uk/2012/09/why-i-will-no-longer-review-or-publish.html.
5. Neuroskeptic, "Fixing science—systems and politics," http://blogs.discovermagazine.com/neuroskeptic/2012/04/14/fixing-science-systems-and-politics/.
6. In my original proposal it was called a registration report, which I soon changed because it sounded like some form of tedious government bureaucracy.
7. See R. Rosenthal, *Experimenter Effects in Behavioral Research* (New York: Appleton-Century-Croft, 1966). When I wrote to Rosenthal in 2015 to inform him that we had finally put his plan in place after nearly 50 years, he said he was pleased that we had implemented his "pipe dream" from the mid-1960s.
8. G. William Walster and T. Anne Cleary, "A proposal for a new editorial policy in the social sciences," *American Statistician* 24, no. 2 (1970): 16–19, http://dx.doi.org/10.1080/00031305.1970.10478884. Similar proposals were later mooted by Robert Newcombe in 1987 and Erick Turner in 2013. See Robert G. Newcombe, "Towards a reduction in publication bias," *BMJ* 295, no. 6599 (1987): 656–59, http://dx.doi.org/10.1136/bmj.295.6599.656, freely available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1257777/pdf/bmjcred00037–0042.pdf; and Erick H. Turner, "Publication bias, with a focus on psychiatry: Causes and solutions," *CNS Drugs* 27, no. 6 (2013): 457–68, http://dx.doi.org/10.1007/s40263–013–0067–9.
9. For my original open letter to *Cortex*, see http://neurochambers.blogspot.co.uk/2012/10/changing-culture-of-scientific.html.
10. I considered the possibility of proposing the idea initially in private and then going public only if it was rejected, but I was concerned that this could reek of sour grapes. Also, this strategy wouldn't have allowed the proposal to benefit from wider peer review.
11. At the time of writing, the Registered Reports subcommittee at *Cortex* includes me, Dr. Rob McIntosh from the University of Edinburgh, Dr. Pia Rotshtein from the University of Birmingham, Professor Klaus Willmes from RWTH Aachen University, and Zoltan Dienes from the University of Sussex.
12. This feature was inspired by a remark from psychologist Hal Pashler, who in commenting in one of Neuroskeptic's posts on study preregistration

(see Neuroskeptic, "Fixing science—systems and politics," http://blogs.discovermagazine.com/neuroskeptic/2012/04/14/fixing-science-systems-and-politics/), said, "reviewers should get to specify some outcome-neutral criteria for publishing the study, e.g., that you do not have a floor effect or ceiling effect, that manipulation checks turn out OK, etc. If you don't do this, then you are asking journals to precommit to publishing studies that fail to offer real tests of hypotheses."

13. For an example of Registered Reports guidelines in full, the reader is directed to the *Cortex* format: http://cdn.elsevier.com/promis_misc/PROMIS%20pub_idt_CORTEX%20Guidelines_RR_29_04_2013.pdf.
14. *Perspectives on Psychological Science* offers a unique twist on Registered Reports, focused on replications and calling for multisite collaborations attempts. It also provides some funding to support these studies. The *Perspectives* initiative was devised independently of Registered Reports by Dan Simons, Alex Holcombe, and others.
15. Chris Chambers, Marcus Munafò, and 83 signatories: "Trust in science would be improved by study pre-registration," http://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration.
16. The main opposition to Registered Reports is exemplified in this response to our *Guardian* article by Sophie Scott, professor of cognitive neuroscience at University College London: https://www.timeshighereducation.com/comment/opinion/pre-registration-would-put-science-in-chains/2005954.article. For a collection of mostly negative reactions assembled by Scott, see also https://sites.google.com/site/speechskscott/SpeakingOut/willpre-registrationofstudiesbegoodforpsychology.
17. When I make this point in seminars I sometimes get asked, "Why aren't *you* proposing an initiative to support exploratory science?" The answer is that we are. An initiative called Exploratory Reports is currently in development at *Cortex*, led by Rob McIntosh.
18. Since February 2014, the Open Scoop Challenge has remained standing and unclaimed: http://software-carpentry.org/blog/2014/02/open-scoop-challenge.html.
19. An additional disincentive for reviewers to scoop is that the "manuscript received" date in the final published Registered Report refers to the initial Stage 1 submission date and so will predate the "manuscript received" date of any standard submission published by a competitor. Therefore, even if a reviewer went ahead and stole the idea and published it first, the original authors could prove they had the idea first.
20. By splitting the review process into two stages and preventing reviewers

from assessing the quality of papers according to the results of hypothesis tests, we also prevent a form of bias by peer reviewers known as CARKing: "critiquing after results are known" (a term coined by Brian Nosek and Daniël Lakens: https://osf.io/vwfk2/). CARKing is a form of motivated reasoning in which a reviewer, objecting to the *results*, raises spurious concerns about the methods in order to prevent publication. For conventional (unregistered) articles, CARKing is impossible to prove—when methods and results are reviewed at the same time, there is no definitive way to distinguish a true methodological objection from CARKing. For Registered Reports, however, any CARKing is obvious and easy to prevent. We see evidence of CARKing if a reviewer approves a submission at Stage 1 but then raises new objections to the same methods after results are presented at Stage 2. While reviewers are free to enter such comments into Stage 2 reviews, the validity of the (already approved) method is not one of assessment criteria; therefore CARKing is not only transparent but impossible.

21. Hint: they almost certainly wouldn't. As a former colleague and eminent neuroscientist once told me: "Never apply for a grant for something you haven't already done."
22. See https://sites.google.com/site/speechskscott/SpeakingOut/willpre-registrationofstudiesbegoodforpsychology.
23. See http://neurochambers.blogspot.co.uk/2012/10/changing-culture-of-scientific.html?showComment=1349772625668#c5836301548034236209.
24. See https://sites.google.com/site/speechskscott/SpeakingOut/willpre-registrationofstudiesbegoodforpsychology.
25. This insult was an extraordinary misfire given that the signatories included, among other luminaries, Dorothy Bishop, Morton Ann Gernsbacher, John Hardy, John Ioannidis, Steven Luck, Barbara Spellman, and Jeremy Wolfe.
26. Some opponents of preregistration suggested that we should have preregistered the Registered Reports initiative (e.g., https://nucambiguous.wordpress.com/2013/07/25/preregistration-a-boring-ass-word-for-a-very-important-proposal/#comment-540). Interestingly, what these critics seemed not to consider is that by advocating preregistration as a way of presumably enhancing the credibility of Registered Reports, they tacitly assume that preregistration does something useful in the first place.
27. Loren K. Mell and Anthony L. Zietman, "Introducing prospective manuscript review to address publication bias," *International Journal of Radiation Oncology, Biology, Physics* 90, no. 4 (2014): 729–732, http://dx.doi.org/10.1016/j.ijrobp.2014.07.052.

28. Robert M. Kaplan and Veronica L. Irvin, "Likelihood of null effects of large NHLBI clinical trials has increased over time," *PLOS ONE* 10, no. 8 (2015): e0132382, http://dx.doi.org/10.1371/journal.pone.0132382.
29. Recent examples at *Cortex* include Jona Sassenhagen and Ina Bornkessel-Schlesewsky, "The P600 as a correlate of ventral attention network reorientation," *Cortex* 66 (2015): A3–A20, http://dx.doi.org/10.1016/j.cortex.2014.12.019; Tim Paris, Jeesun Kim, and Chris Davis, "Using EEG and stimulus context to probe the modelling of auditory-visual speech," *Cortex* (2015), http://dx.doi.org/10.1016/j.cortex.2015.03.010. These and more are showcased in a virtual special issue of Registered Reports at *Cortex*: http://www.journals.elsevier.com/cortex/virtual-special-issues/virtual-special-issue-registered-reports. See also the special issue on Registered Reports of replications in *Social Psychology*: http://econtent.hogrefe.com/toc/zsp/45/3.
30. For a full list of currently participating journals and guidelines, see https://cos.io/rr/.
31. http://www.acmedsci.ac.uk/policy/policy-projects/reproducibility-and-reliability-of-biomedical-research/.
32. For more information on the TOP guidelines, see https://cos.io/top/.
33. I am the current handling editor for Registered Reports at Royal Society Open Science, https://blogs.royalsociety.org/publishing/registered-reports/. Even though the physical sciences suffer less from questionable research practices and researcher bias, they are nevertheless subject to publication bias and therefore stand to benefit from Registered Reports.
34. The freely available Open Science Framework (http://osf.io) provides a versatile registry for many different kinds of research, including psychology. Other options include Figshare, the American Economic Association's registry for randomized trials (https://www.socialscienceregistry.org/), and for studies with health implications, researchers can also use the ISRCTN registry (http://www.isrctn.com/) or http://clinicaltrials.gov.
35. Leif Nelson, "Preregistration: Not just for the empiro-zealots," http://datacolada.org/12.
36. Sylvain Mathieu, Isabelle Boutron, David Moher, Douglas G. Altman, and Philippe Ravaud, "Comparison of registered and published primary outcomes in randomized controlled trials," *JAMA* 302, no. 9 (2009): 977–84, http://dx.doi.org/10.1001/jama.2009.1242.
37. Sreeram Ramagopalan, Andrew P. Skingsley, Lahiru Handunnetthi, Michelle Klingel, Daniel Magnus, Julia Pakpoor, and Ben Goldacre, "Prevalence of primary outcome changes in clinical trials registered on ClinicalTrials.gov: A cross-sectional study," *F1000Research* 3 (2014), http://dx.doi.org/10.12688%2Ff1000research.3784.1.

38. Sreeram V. Ramagopalan, Andrew P. Skingsley, Lahiru Handunnetthi, Daniel Magnus, Michelle Klingel, Julia Pakpoor, and Ben Goldacre, "Funding source and primary outcome changes in clinical trials registered on ClinicalTrials.gov are associated with the reporting of a statistically significant primary outcome: A cross-sectional study," *F1000Research* 4 (2015), http://dx.doi.org/10.12688%2Ff1000research.6312.2
39. http://compare-trials.org/#.
40. Annie Franco, Neil Malhotra, and Gabor Simonovits, "Underreporting in psychology experiments: Evidence from a study registry," *Social Psychological and Personality Science* 7, no. 1 (2016): 8–12, http://dx.doi.org/10.1177/1948550615598377.
41. The use of standardized templates could help increase the precision of unreviewed preregistration. A new tool developed by Uri Simonsohn and Joe Simmons shows promise in this regard: https://aspredicted.org/.
42. Adam Marcus and Ivan Oransky, "Time for a reproducibility index," http://www.labtimes.org/labtimes/ranking/dont/2013_04.lasso.
43. A complementary idea for a "Replication Tracker" has been proposed by Josh Hartshorne and Adena Schachner, http://dx.doi.org/10.3389/fncom.2012.00008.
44. At least not without a sophisticated artificial intelligence that is beyond our current capabilities.
45. For details, see http://www.iupsys.net/about/members/national-members/index.html.
46. One avenue for producing a Reproducibility Index could be to develop it conjunction with the Curate Science project led by Etienne LeBel, http://curatescience.org/. This impressive new initiative tracks the replicability of published results. For an introduction, see http://replicationnetwork.com/2015/10/21/lebel-introducing-curatescience-org/
47. Leonard P. Freedman, Iain M. Cockburn, and Timothy S. Simcoe, "The economics of reproducibility in preclinical research," *PLOS Biol* 13, no. 6 (2015): e1002165, http://dx.doi.org/10.1371/journal.pbio.1002165.
48. For the full study, see Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science* 349, no. 6251 (2015): aac4716, http://dx.doi.org/10.1126/science.aac4716. Alex Etz and Joachim Vandekerckhove have provided a Bayesian analysis of these results; they find that about 20 percent of the attempted replications provided strong evidence for a replication failure and 25 percent provided strong evidence for replication success, with the remainder occupying a middle ground of moderate or inconclusive evidence either way: Alexander Etz and Joachim Vandekerckhove, "A Bayesian perspective on the reproducibility project: Psychology,"

PLOS ONE 11, no. 2 (2016): e0149794, http://dx.doi.org/10.1371/journal.pone.0149794. A lucid summary of these findings written for a more general audience can be found on Alex Etz's blog: http://alexanderetz.com/2015/08/30/the-bayesian-reproducibility-project/.

49. The Registered Reports funding scheme was proposed in 2016 by Marcus Munafò and myself and is currently under consideration by a number of journals and funders.
50. Kevin M. Williams, Craig Nathanson, and Delroy L. Paulhus, "Identifying and profiling scholastic cheaters: Their personality, cognitive ability, and motivation," *Journal of Experimental Psychology: Applied* 16, no. 3 (2010): 293, http://dx.doi.org/10.1037/a0020773. The article can be freely read at https://www.apa.org/pubs/journals/releases/xap163293.pdf.
51. Ibid., 305.
52. "Diederik Stapel settles with Dutch prosecutors, won't face jail time," http://retractionwatch.com/2013/06/28/diederik-stapel-settles-with-dutch-prosectors-wont-face-jail-time/.
53. "Former University of Queensland professor Bruce Murdoch charged over alleged fake Parkinson's research," http://www.abc.net.au/news/2014–12–12/university-of-queensland-professor-on-fraud-charges/5964476.
54. Dea Clark, "Former UQ academic Dr Caroline Barwood granted bail over fraud charges," http://www.abc.net.au/news/2014–11–06/former-uq-academic-granted-bail-over-fraud-charges/5871436.
55. There will always be some disadvantages for those who are close to the suspected fraudster, such as the retraction of coauthored papers. It would be unrealistic to believe that we can ever make the act of whistle-blowing completely harmless for the whistle-blower, but we can certainly do more to minimize the damage.
56. For further discussion, see Chris Chambers and Petroc Sumner, "Replication is the only solution to scientific fraud," http://www.theguardian.com/commentisfree/2012/sep/14/solution-scientific-fraud-replication.
57. For a list of RoMEO green journals, see http://www.sherpa.ac.uk/romeo/browse.php?colour=green.
58. Björn Brembs, Katherine Button, and Marcus Munafò, "Deep impact: Unintended consequences of journal rank," *Frontiers in Human Neuroscience* 7 (2013): 291, http://dx.doi.org/10.3389/fnhum.2013.00291.
59. Sam Schwarzkopf, "Revolutionise the publication process," http://neuroneurotic.net/2015/07/17/revolutionise-the-publication-process/; Dorothy Bishop, "Will traditional science journal disappear?," http://www.theguardian.com/science/head-quarters/2015/may/12/will-traditional-science-journals-disappear. Another related idea is the Episciences project,

which preserves the journal model but seeks to conduct peer review and publishing entirely with so-called overlay journals that operate alongside open archives such as arXiv. See http://www.episciences.org/.

60. There are already gentle moves in this direction. Some journals now require authors to stipulate their contribution to various aspects of a paper. For instance, *PNAS* now requires authors to state their contribution to research design, performing the research, contribution of new reagents/analytic tools, data analysis, and writing of the paper.
61. Although an alphabetic model is preferable, in the transition between a contributorship model and the traditional authorship model, the first author position could be retained, with all subsequent authors listed alphabetically and contributorship details noted for all authors.

    For an insightful discussion of contributorship models of attribution, see Amy Brand, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott, "Beyond authorship: Attribution, contribution, collaboration, and credit," *Learned Publishing* 28, no. 2 (2015): 151–55, which can be freely downloaded from http://openscholar.mit.edu/sites/default/files/dept/files/lpub28–2_151–155.pdf.
62. The danger of false precision could be addressed by replacing point value percent contributions with either a percent range or a descriptive scale. Various methods could be trialed.
63. https://twitter.com/david_colquhoun/status/664949963890278400.
64. Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn, "A 21 word solution," available at SSRN 2160588 (2012), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2160588.
65. See http://opennessinitiative.org/.
66. See www.sherpa.ac.uk/romeo/.
67. In addition to the data partner scheme proposed by Richard and Candice Morey in chapter 4 (see http://bayesfactor.blogspot.no/2015/11/habits-and-open-data-helping-students.html), Kia Nobre's lab at the University of Oxford has recently proposed a similar "frenemy system," https://sites.google.com/site/todorovicana/musings/crisis_report.
68. http://www.ascb.org/dora/.
69. For details about the TOP guidelines, see https://cos.io/top/.
70. For details about Registered Reports, see https://cos.io/rr/.
71. For details about the Badges initiative, see https://osf.io/tvyxz/wiki/home/.

INDEX

Page numbers followed by *f* indicate a figure.