Open Minded

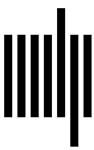
Searching for Truth about the Unconscious Mind

By: Ben R. Newell, David R. Shanks

Citation:

Open Minded: Searching for Truth about the Unconscious Mind
By: Ben R. Newell, David R. Shanks
DOI: 10.7551/mitpress/14922.001.0001
ISBN (electronic): 9780262375375
Publisher: The MIT Press
Published: 2023

The open access edition of this book was made possible by generous funding and support from MIT Press Direct to Open



The MIT Press

10 Research Reformed

It is plain that selective reporting, *p*-hacking, and other common practices can lead to gross distortion of the scientific record, with published results commonly presenting a false picture of reality. An obvious question at this point is why the practice of replication (repeating a previous study's methods to obtain new data and see if similar results are obtained) does not rapidly weed out spurious findings. As we noted in the previous chapter, anyone trying to market a nonfunctional new type of battery would instantly be found out. Incorrect scientific or technological developments do not last long under the glare of immediate feedback. When chemists Martin Fleischmann and Stanley Pons announced in 1989 that they had produced cold fusion, the prospect of almost limitless clean energy galvanized the public and media. But within a few weeks, after many independent teams had failed to confirm Fleischmann and Pons's findings, the *New York Times* declared that cold fusion was dead. If money priming is not a real phenomenon, then why didn't failed replications immediately reveal this?

Replication and Registration

At various points in this book, we have discussed examples of replication failures. In chapter 3, we described two famous psychological experiments (walking slowly and smiling through your teeth), neither of which proved to be replicable. In chapter 4, we briefly reviewed another one, on incidental anchoring (people don't pay more at restaurants with high numbers in their names). Because the results of individual studies can be incorrect due to flaws or random error, replication is fundamental to confirming the validity of a scientific claim. Indeed we could go further and imagine a world in

which every published result in science was ignored by other scientists, as well as by the media, until a successful replication was reported. The truly terrible consequences of the false link between the measles, mumps, and rubella vaccine and autism would never have happened, for example.[1]

Unfortunately, direct replication plays a tiny role in most scientific fields. Estimates consistently suggest that only around 1 percent of all psychological research is ever replicated, a state of affairs that is almost universally recognized as needing to change.[2] Part of the problem is that, as we discuss later, science is a competitive field, and researchers often think that they will receive little reward for investing time and effort into "merely" reproducing a result that is already known. After all, the glory goes to the person who first made the discovery, not the unimaginative drudge whose contribution is simply and boringly to confirm it. But in light of what we now know about publication bias and *p*-hacking, researchers are starting to undertake more and more replications, particularly of eye-catching results.

Money priming is one such result, and a major replication study led by Richard Klein of the University of Florida sought to reproduce it in a very large-scale, multilab project.[3] Thirty-six teams from around the world agreed to participate, each running the same battery of tests designed to generate thirteen well-known effects, including money priming. With a near-identical procedure to one of the original experiments demonstrating the effect, participants began by answering some demographic questions via computer. For some of them, the background was a faint picture of $100 bills, while for others, it was a blurred version of the picture in which the bills could not be identified as such. Then participants answered questions regarding their attitude to the fairness and legitimacy of the prevailing social system. A typical item was, "Everyone has a fair shot at wealth and happiness," rated from "strongly disagree" to "strongly agree."

Figure 10.1 reproduces the funnel plot from the previous chapter, but now adds the thirty-six individual results from this multilab replication project, as well as those from another replication effort, indicated by open triangles.[4] A couple of things are immediately apparent. First, these replication results are generally higher in the figure than the original studies. This means that they yield more precise estimates (they have smaller standard errors), which in turn comes from the fact that they employed substantially larger samples than the original experiments: while the original experiments tested a median sample of only 66 participants, the replications had a median

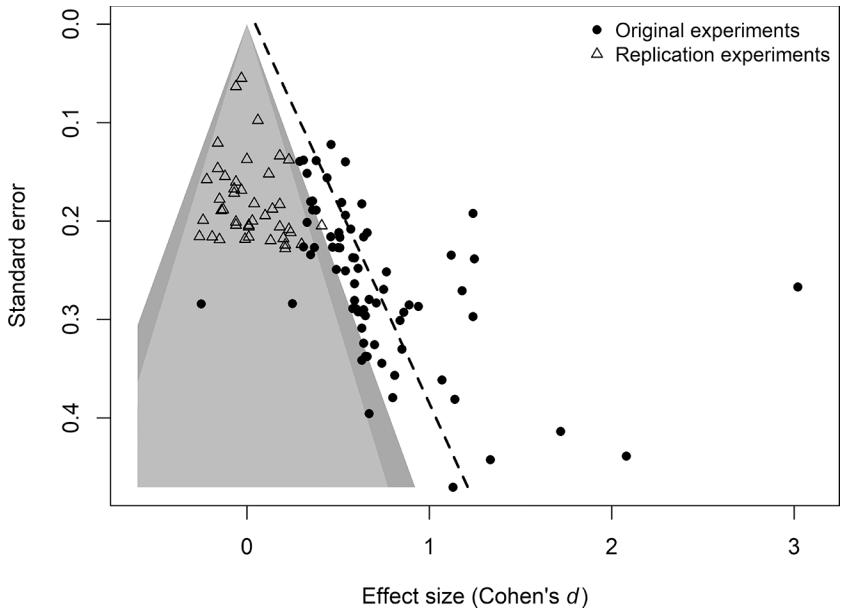


Figure 10.1
This figure reproduces figure 9.1 but adds the results of preregistered replication experiments (open triangles). The vertical axis represents each experiment's precision, while on the horizontal axis is depicted the outcome of each experiment, measuring the size of the money priming effect in the standardized measure, Cohen's *d*. Experiments falling to the right of the gray funnel have statistically significant ($p < .05$) results, while those falling inside the funnel have nonsignificant ($p > .05$) results (the dark gray region indicates marginally significant effects falling between $p = .05$ and $p = .10$). The dotted trend line is fitted to the original experiments only.

sample of 110. This is still not very large by the standards, say, of medical trials, but at least it is a step in the right direction. The second obvious aspect of the replication results is that they fall symmetrically within the funnel. Unlike the asymmetry of the distribution of original effects, the replications show no tendency for a relationship between precision and effect size. Finally, and most important, they cluster around an effect size of 0—that is, no overall difference in the attitudes to the prevailing social system of participants primed with money compared to those not primed. Indeed there is almost no overlap in the effect sizes of the original and replication experiments. Despite there being literally hundreds of studies appearing to obtain money priming effects, a near-exact replication project failed completely to detect an effect.

How can this be? The replications vindicate the conclusions of the funnel plot discussion in the previous chapter and bolster the inference that many of the apparently successful demonstrations of money priming are false positives—results appearing to find an effect that is not real in the population—resulting from *p*-hacking or good fortune (the researchers ran many studies, and the published ones are those that by chance found a statistically significant effect). The many "missing" studies in the funnel plot are the telltale clue attesting to this. But one might ask why the replication studies are any more believable than the original ones. Don't we simply have a case here in which one set of studies disagrees with another set? After all, the fact that Klein's replication project was conducted after many of the original studies were reported is neither here nor there: the original studies fail to replicate the null findings of the replications to just the same extent that the replications fail to replicate the positive findings of the original experiments.

To see how implausible this interpretation is, remember that thirty-six independent teams contributed to the replication, and these teams had a diverse set of expectations about what they would find, some believing the effect would be found and others not. If special expertise and care or expectations are needed to obtain the effect, then at least a few of the teams should have found a positive money-priming effect, but in fact only one team did—just as would be expected by chance bearing in mind that $p < .05$ implies a lucky positive result once in every twenty or so attempts.

But the strongest reason to believe the replication findings over the original ones is that Klein and his colleagues preregistered their entire study. Before collecting any data, they carefully described exactly how their study would proceed and how the data would be handled and analyzed, effectively tying their own hands to prevent any possibility of later *p*-hacking. As promised in the preregistration, the data were not examined until all testing had been completed. The preregistration was uploaded to a public repository together with the program for the experiment itself in advance, so anyone can go back and check that they did exactly what they said they would do. These features are in stark contrast to standard experimental practice. For each of the "biased" studies in the funnel plot (the original experiments), we simply do not know whether multiple analyses were run and only the significant ones published; whether participants were added or removed after running initial, exploratory analyses; or whether these studies are only a subset of

all the studies ever conducted by those teams (and we can't know what those other unpublished studies would look like). In preregistered studies, in contrast, what you see is all there is.

Preregistration is rapidly becoming a crucial method for boosting the credibility of research, going a long way to eliminating many of the evils discussed previously.[5] While *p*-hacking is the most obvious one, others are eliminated as well. Because the preregistration describes the experimental hypothesis in detail, the researcher's ability to indulge in flexible retrospective HARKing (reinterpreting a surprising result as if it were predicted all along) is severely curtailed. Publication bias is also appreciably less likely, not only because the preregistration is published in the sense of being a publicly accessible document, but also because as long as the study was executed in accordance with the stated plan, its findings are likely to be a contribution to the academic literature: if it replicates the finding it was attempting to repeat, then it's a valuable confirmation of that finding, whereas if it fails to replicate the earlier result, that itself is important knowledge.

Preregistration powerfully emphasizes the crucial distinction between two forms of research endeavor briefly mentioned previously: exploration versus confirmation. Exploratory research is what we all have in mind when we think of a scientist working at a laboratory bench, trying to make a discovery, solve a problem, or build a new device. Exploration is unquestionably the engine of scientific and technological advancement, as well as being the main yardstick against which scientists themselves are judged. But confirmatory work—carefully seeking to validate previous claims and findings—is just as important. It is an essential tool for us to separate out true findings from all the *p*-hacked false ones. Indeed some have argued that psychological research in general should move toward a model in which research publications comprise initial exploratory studies followed by large-scale confirmatory ones.[6] But what does an ideal confirmatory study look like?

In medical research, it has been compulsory for many years to preregister clinical trials before conducting them. For example, ClinicalTrials.gov, established in 2000, is a repository of, to date, about 400,000 trials. Laws mandating the registration of trials involving drugs or devices have been passed in both the United States and the European Union. This sounds like an excellent mechanism to decrease the chance that the public will be exposed to treatments or drugs that are in fact ineffective. Surely the researchers conducting the trial cannot *p*-hack the results in order to gain a statistically significant

result if their hands are tied by their preregistered commitments about how they would conduct and analyze the trial? And indeed there is evidence that clinical trials are becoming less and less successful.[7] This may sound like bad news, but in an important sense, it's the exact opposite. We have argued that much of the published scientific literature comprises false-positive results, either wrestled out of unpromising data by expert *p*-hackers or simply the lucky survivors of the Darwinian selection process that diverts successful studies into scholarly journals and unsuccessful ones into the file drawer. If this is even remotely correct, then we should expect any mechanism that suppresses *p*-hacking and publication bias to decrease the number of false positives in the scientific record. So the fact that fewer and fewer trials in some domains are succeeding may, paradoxically, be a good sign.

Unfortunately this form of preregistration does not provide any iron-clad guarantee that research practices will improve. It fails to protect against publication bias because the researcher may choose not to submit or a journal may choose not to publish the results if they are messy or negative. Moreover, and somewhat amazingly, analyses show that researchers engage in widespread *p*-hacking even when their public preregistered methods descriptions make it easy for anyone to spot the *p*-hacking. A prime example of this is the switching of a trial's designated primary outcome. Imagine that a trial is being run to measure the efficacy of a new medicine in treating headaches. As part of the preregistration, the researcher may announce that the number of headaches per week is the crucial outcome measure, the one on which the trial's success stands or falls. This measure either shows a statistically significant decline, in which case the trial has been successful, or it doesn't. Later, the trial is published in a scholarly journal, but now the key outcome measure that the article analyzes is headache duration, not number. The researcher has switched outcomes between preregistration and publication of the results. Obviously a likely explanation is *p*-hacking: the effect was statistically significant on the duration but not the number measure, so the researcher switched them in order to get the article published. The scale of these reporting switches is alarming, and consistent with the *p*-hacking explanation: when outcomes are switched, they overwhelmingly tend to be in favor of achieving statistical significance.[8]

These switches also serve to highlight (if we needed further evidence of this) that the peer review process falls far short of providing a guarantee

that published research is credible and maintains high standards of research probity and rigor. One might hope that reviewers would immediately spot these switches and other deviations from the preregistration and reject the paper for publication, but this rarely happens. Peer reviewers are unrewarded for their work and have little incentive to spend undue amounts of time cross-checking a manuscript against a preregistration, and indeed there is concrete evidence that they rarely do so.[9] Preregistration is a step in the right direction and at least means that *p*-hacking becomes visible to anyone wishing to compare a preregistration against published results, but it is not a cast-iron method of boosting research credibility.[10]

A stronger form of preregistration is beginning to gain a foothold and may prove to be the ideal format for conducting confirmatory research. In so-called *registered reports*, the researcher describes in complete detail how she plans to carry out a study or experiment, as well as the hypothesis being tested, the primary outcome measure, the data analysis method, and so on.[11] But instead of simply posting this description on a time-stamped public repository and then proceeding to collect the data, as would be the case under standard preregistration, she instead submits the description to a journal for evaluation. The journal asks reviewers to assess the described study for its rigor (for instance: Will its sample size be adequate? Is the method appropriate to test the hypothesis?) and likely contribution to the field, and then if it is judged of sufficient quality guarantees to publish it once the study has been completed, *regardless of the results*. The journal is in effect making a results-blind decision about the work that places all the emphasis on the rationale and methodological rigor of the study and none on the results. The results will be what they will be. In the process of approving the final article for publication, the journal reviewers are asked to check that the researcher conducted the study according to the preregistration description, has explicitly noted any deviations, and clearly flags any new analyses that were not preplanned as exploratory ones not to be confused with the primary confirmatory analyses.

It's easy to see that registered reports of this form, which are now solicited and published in many journals, provide a high level of protection against selective publication, *p*-hacking, HARKing, and so on. The results are published regardless of what they reveal, hence protecting against the selective nonreporting of statistically nonsignificant results. The researcher

precommits to the analysis, eliminating the scope for *p*-hacking. And because the hypothesis is stated in advance, there is minimal scope for seeing the results and then going back to change the purpose of the study.

Indeed it is now becoming apparent that the quality of research published in registered reports is appreciably higher than in standard peer-reviewed journal articles. In a recent project led by Courtney Soderberg from the Center for Open Science, a nonprofit organization based in Charlottesville, Virginia, over three hundred experts were recruited as assessors.[12] Each was given a deidentified and lightly redacted registered report (thus reducing the likelihood that the assessor realized that it was a registered report) as well as a carefully selected and matched standard journal article to evaluate. Across nineteen evaluation criteria, the registered reports scored higher on all dimensions. Their methods and analyses were rated as more rigorous, they were judged more novel and creative, the quality of the discussion of the findings was judged better, and so on. As a means of enhancing research quality (as well as boosting public trust in science), it would not be an exaggeration to suggest that future generations will look back at the registered report format as one of the most significant methodological developments in the history of science.

There is a clear further test of the idea that registered reports provide protection against publication bias and *p*-hacking: they should yield positive findings much less frequently than standard nonregistered publications. Put the other way around, null results should be much more common in registered reports than elsewhere. This issue, which we've already touched on in relation to unpublished studies and clinical trials, is a key indicator of the credibility of the scientific literature. We know that the vast majority of published studies report positive—in other words, statistically significant—findings. Across social and behavioral research, estimates of the proportion of null results vary but are generally in the range of 5 to 20 percent.[13] What do we find when we look at the rate of null findings in properly preregistered experiments? It is startlingly higher. Although the researchers carrying out these registered experiments have framed plausible hypotheses and tested large samples of participants, their carefully and publicly preplanned experiments are successful only in a minority of cases in finding a meaningful effect or group difference. In at least half of all cases, they yield a null result.[14] These findings tell us loud and clear that the high rate of positive

findings in the normal scientific literature (at least 80 percent) cannot be a true and unbiased reflection of the world.

The money-priming literature provides a perfect illustration of this difference. In a comprehensive meta-analysis of all available studies,[15] including 47 preregistered tests and 189 standard non-preregistered ones, 62 percent of all standard studies obtained positive results but only 11 percent of the preregistered ones did. This same pattern can be seen in the funnel plot shown in figure 10.1. The open triangles all come from preregistered studies and find effects close to 0, whereas the other points reflecting standard experiments are much more likely to indicate positive effects. Recall that the gray area in the funnel depicts all effects that are statistically nonsignificant.

There is one further consequence of the greater prevalence of null results in preregistered experiments, combined with the growing frequency of such experiments: we should see many effect sizes dwindling over time. If early reports are biased by *p*-hacking and publication bias, whereas later preregistered studies ameliorate these biases, then observed effects should become smaller and smaller, and this will be true whether they are genuine effects or completely spurious. In the former case, the effect size will eventually converge on the true positive estimate. This "decline effect" is what has happened with studies on cognitive-behavior therapy (CBT) for depression, for instance, with the observed efficacy of this type of treatment slowly dwindling over the past forty or so years.[16] Despite this, the most up-to-date estimates still show it to be quite effective.

In the case of truly spurious effects, we would expect the estimated effect size to eventually converge on 0. The final nail in the coffin of money priming is that studies show exactly this pattern. Figure 10.2 graphs the effect sizes of money-priming tests, including both published and unpublished studies, some preregistered and some not, across time. Although the data go up to only 2017, it is clear that the effect has been declining steadily since the original studies in 2005 and 2006. The best estimate of the outcome of a money-priming study conducted after 2018 is very close to zero. After all this huge effort studying an eye-catching way of unconsciously nudging people's behavior and the vast amount of journal space devoted to it (not to mention the many taxpayer-funded research grants), we find that the effect proves to be no more real than the telepathy, clairvoyance, and extrasensory perception effects first studied experimentally in the 1940s by J. B. Rhine.

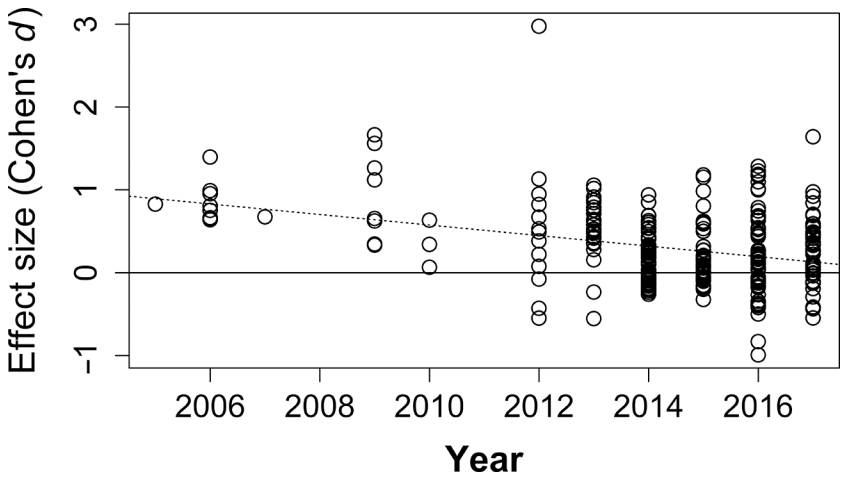


Figure 10.2
This figure depicts, across each year since the original money-priming experiments were published, the effect size (in Cohen's *d* units) of every test that has been conducted (the data were compiled by Paul Lodder). Positive effect sizes represent effects in the direction expected by the money-priming hypothesis (for instance, that subtle reminders of money cause people to work harder on difficult tasks). Remarkably, up until 2012, every study yielded an effect in the predicted direction. Since then more and more replications yielding null results, including preregistered studies, have been conducted, including those from the multilab project led by Richard Klein that contribute thirty-six of the data points scattered close to 0 in 2014. The dotted trend line shows that the overall effect has been steadily declining since it was first reported.

The Power of Myths

There is another reason that eye-catching claims in psychology, including ones like money priming that concern unconscious mental processes, often maintain high prominence and uncritical acceptance long after they have been discredited. It is because we are strongly persuaded by story lines and myths that make sense of the complex and confusing world around us. Saying that many aspects of our behavior arise through unconscious influences seems a simple and parsimonious way of explaining actions that would be fiendishly hard to rationalize any other way. If you were asked to explain why you worked unusually hard on a given task—cleaning the house one day, say—you would probably struggle to come up with a convincing

explanation in terms of the conscious thoughts and motivations that went through your mind at the time. But saying that you were influenced by numerous factors of which you were largely unaware—such as your eyes fleetingly glancing at a banknote lying on the table—provides a compelling story. You don't even have to enumerate all the factors that influenced you. That's part of the beauty of the explanation—that these factors were unconscious and hence you can't report them.[17]

And this storytelling is part of science itself. Take the case of citations to research that has been strongly contradicted in replication efforts. When scientists write up their research for journal publication, they introduce their project and its purpose by reference to previous relevant work. This introduction section, together with the discussion presented at the end of the article once the results have been described, tries to present a narrative that makes sense of the results and fits them into a larger perspective on the topic. In citing relevant previous research, one would expect a fair-minded approach in which the strengths and weaknesses of the earlier work are evaluated from a neutral perspective, regardless of whether the cited work fits in with or runs counter to the author's own perspective. When we look at citation patterns, we soon see that this assumption is a gross idealization. Particularly striking is that high-profile original studies continue to be cited at high rates even if they've been strongly contradicted by subsequent replications. Money priming provides a clear illustration.[18] Despite the fact that Klein's large-scale, multilab project completely failed to replicate money priming, researchers continue to cite the original research by Vohs and her colleagues almost as if the replications didn't exist. In the five years following the publication of Klein's failed replication, the number of annual citations of one of the key original reports continued unabated. One might imagine that later researchers were citing the original report in the context of discussion about its unreplicability, but this was not the case. The vast majority of these citations were favorable, discussed money priming as if there was no problem with it, and did not cite the Klein article. Moreover, across several case studies of this type, even in those instances where authors did cite both the original research and the replication failure, they often provided no explicit justification for their favorable assessment of the original research.

Scientific textbooks, one of the main ways in which knowledge in a discipline is transmitted to new members of the discipline (that is, students) and interested laypeople, provides another illustration of how myths can lead

to distorted evaluation of research. Numerous case studies document the myths that textbooks help to sustain. A fascinating and instructive one concerns what is probably the most famous experiment in all of psychology, the Stanford prison experiment. In August 1971 Philip Zimbardo assigned students by a coin toss to the role of "guards" or "prisoners" in a mocked-up prison in the basement of the Stanford University Psychology Department, with Zimbardo playing the role of prison superintendent. The experiment had to be shut down after six days because the students adopted their roles rather too convincingly. The guards started to commit acts of psychological torture on the prisoners, some of whom accepted their roles as victims of abuse. The experiment is widely taken as providing evidence that the context (including the social roles placed on us) plays a far greater role in determining human behavior than individual personal dispositions such as our particular personality attributes.

But this is little more than a story.[19] From a scientific point of view, the Stanford prison experiment comes nowhere close to demonstrating the power of social roles. The guards didn't act as they did because of their roles as guards, but because Zimbardo effectively instructed and guided them in how he expected them to behave. Subsequent reports from those who took part make this abundantly clear. Carlo Prescott, an ex-convict who served as chief consultant to Zimbardo on real prisons, later said that "Zimbardo began with a preformed blockbuster conclusion and designed an experiment to 'prove' that conclusion." John Mark, one of the guards in the experiment, commented that Zimbardo "knew what he wanted and then tried to shape the experiment. . . . He wanted to be able to say that college students, people from middle-class backgrounds—people will turn on each other just because they're given a role and given power." Most striking, in later attempts to replicate the experiment in which the guards were not directly instructed to abuse the prisoners, findings quite different (though no less interesting) from those of the Stanford prison experiment emerged.[20]

Despite the fact that the experiment falls far short of demonstrating its primary claimed conclusion, scientific textbooks continue to spread the conventional story about its significance. A detailed survey of seven contemporary social psychology textbooks written by experts in the field that included discussion of the Stanford prison experiment found that only two provided anything approaching a balanced discussion of the criticisms leveled against it. Closer to home, the same biased reporting is evident in discussions about

the unconscious mind. In chapter 5 we described the implicit association test (IAT), a workhorse tool for purportedly measuring unconscious racial and other forms of bias, and some of the many criticisms leveled against this tool. A major concern is the paucity of evidence (despite many efforts to find such evidence) that IAT scores predict observable real-world behaviors indicative of bias. In an analysis of the way the IAT is discussed in seventeen introductory psychology textbooks, only two mentioned the dubious record of the IAT as a predictive tool.[21] It seems far more acceptable to textbook authors to tell a largely mythical story about psychological research than to give a more nuanced (but arguably more truthful) assessment. Discussing problems with a piece of research might muddle the story and create confusion in readers' minds. A good story, in contrast, may engage students and help to sell textbooks, but at the cost of misrepresenting reality.

Science does a poor job of correcting itself when initial eye-catching findings are later found to be either partially or wholly incorrect or are for some other reason discredited. Every year there continue to be numerous favorable citations to the research of the Dutch social psychologist Diederik Stapel, despite the fact that he fabricated data for his experiments (and admitted as much—as we saw in chapter 8). His studies have been formally retracted by the journals in which they were originally published but remain accessible. Despite the retractions, nontrivial numbers of scientists continue to discuss the findings of his research as if they have never been challenged.[22] Often this is presumably a consequence of lazy cutting-and-pasting when making a minor point that is not central to the scientist's research report, but it nonetheless highlights the fact that science itself struggles to ensure a balanced assessment of evidence, even in the most extreme and incontrovertible cases.

The Scientific Ecosystem

These problems with the ways in which scientists conduct their research are compounded by an ecosystem that encourages behaviors that are at variance with the pure pursuit of truth. Science is a competitive field. Scientists are employed by institutions that compete for students and prestige, they apply to competitive grant funding agencies to sponsor their research, and they submit their findings to journals that compete for subscribers and citations (when research articles include an earlier publication in their bibliographies, a standard currency for measuring the influence of the cited

publication). Top scientists are headhunted at vast expense. At all of these stages, incentives are created that can pull the researcher away from the neutral pursuit of the truth.

If researchers are incentivized to try to maximize the number of publications they generate or the number of citations their research receives, then as sure as night follows day, they will modify their behavior to achieve these goals, even if it's at the expense of producing high-quality research. Goodhart's maxim tell us that when a measure becomes a target, it ceases to be a good target, and this is manifestly the case in the academic universe in which researchers are rewarded for the number of journal articles they produce and the citations they receive.[23]

One particularly stark but simple illustration that the quantity/quality balance is awry in much of psychological research is provided by surveys of the sizes of the samples researchers employ to test hypotheses in their experiments. This is a basic feature of research. After framing the hypothesis that she wishes to test and the measures and manipulations that will be employed to test it, the researcher must make decisions about the participants who will be tested—their age, characteristics, and, most important for this discussion, how many. It has been known for decades that these sample sizes tend to be too small. Imagine you're a researcher interested in measuring the effect of a particular intervention, say the effect of CBT on the symptoms of depression in primary care. In a standard randomized control trial (RCT), you might administer CBT to one group and a control or placebo treatment to another group. But how many people should be included in each group?

Past meta-analyses on this topic have found that the effect of CBT on depression—one of the best-established nonpharmacological treatments there is—has an effect size of about 0.2 in Cohen's *d* units.[24] Remember that the effect size for the male-female difference in height is about ten times this value, so against this benchmark, the beneficial effects of CBT are quite small, though of course when administered across many patients, this nonetheless amounts to a very meaningful therapeutic benefit. But our question is about statistical "power": How many people should the researcher include in each group of her RCT in order to be reasonably confident (say, 80 percent confident, which is the standard level adopted) of obtaining a statistically significant difference in measured depression symptoms between the two groups? The answer is that about 600 people are needed in total, assuming half are allocated to each group. This is a very large sample; the study would

likely take many months to run and be a considerable time, money, and effort commitment for the researcher.

In reality, studies on the effects of CBT on depression in primary care, of which there have been over 30, have an average sample of about 160, vastly smaller than the sufficient size. Indeed this figure is in line with wider surveys of published research in psychology, suggesting that the average sample size is close to around 100 to 200,[25] and even this figure may wildly exaggerate the average for experimental psychological research, including studies on unconscious mental processes.[26] What this means is that researchers are often running experiments that are too weak to observe effects, even if those effects really exist. It may be the case that typical effects studied in psychological research in the field and laboratory are generally slightly bigger than that of CBT on depression, averaging a Cohen's *d* of, say, 0.4, but the sample sizes researchers use are still too small and are increasing at a glacial rate, if at all.[27]

Why do scientists tend to underpower their studies? The answer is not hard to discern: running smaller studies takes less time and resources and hence enables more articles to be published, leading to faster promotion, a bigger reputation, and so on. One might wonder what the point is of running an experiment with an inadequate sample size. Surely doing so raises the risk that the experiment will yield a nonsignificant, unpublishable result. This is where *p*-hacking comes to the rescue. Switching the outcome variable, for example, may exchange a statistically nonsignificant and unpublishable result for a significant and publishable one. And running underpowered studies doesn't simply increase the chances of wrongly obtaining null results; it also increases the likelihood that those studies that do, by good fortune, yield statistically significant results are false positives. As power decreases in a set of experiments, the ratio of false to true positives increases. It is not hard to see how inadequate sample sizes, resulting from the inherent pressures of the scientific ecosystem, can contribute to the creation of literatures like the money-priming one.

Individual researchers conducting their studies either independently or in collaborative groups represent one point on the pipeline for the generation of published articles. At other points along this pipeline, there are further incentives that can undermine the smooth and unbiased pursuit of truth. Even before investigators begin to collect data, funding agencies and industrial partners decide which projects to support, and it is well known that

the funding source can bias the outcomes of the research. Large-scale meta-analyses reveal, for example, that drug trials funded by pharmaceutical companies yield more favorable outcomes than ones funded from other sources such as national research agencies.[28]

Closer to the end of the research pipeline are scholarly journals whose behavior also establishes perverse incentives that are often not aligned with the pursuit of truth. In the past, there tended to be a degree of commercial separation between a journal's publisher and its editorial team. In many cases, a journal would be strongly affiliated to a non-profit-making learned society whose purpose is to use journal revenue to fund researchers—particularly early-career scientists—working in its particular field, run research workshops and conferences, provide travel grants, and so on. Hence, the model involved universities paying a journal subscription to the publisher and the publisher handing over an agreed annual amount to the learned society for the privilege of having its badge of esteem on the journal. The society would provide the entire editorial team for the journal, deciding on its overall policy and making decisions on each manuscript submitted to it for evaluation.

It is easy to imagine that in such a model, editors have very little at stake other than the preservation of academic rigor, in whether any submitted manuscript is published. Editors, who by day are typically university employees, receive no remuneration for their work and certainly do not stand to gain financially from the publication of submitted manuscripts. But the idea that journals and their editors are disinterested referees solely concerned with maintaining scientific rigor is little more than an idealization. The reality is that even among reputable journals, the scientific ecosystem rewards behaviors that are not necessarily well aligned with the production of high-quality research. There is an understood hierarchy of journal prestige, with journals like *Science* and *Nature* at the very top. It can be a career-changing event for a young scientist to publish an article in one of these—as the fraudulent psychologist Diederik Stapel did in 2011. Usually this prestige is quantified by bibliometric indicators such as the journal's impact factor, which measures how frequently articles in that journal are cited by other researchers across a one-year period following publication. But there is no evidence that the research these journals publish is of higher quality than research published in more modest journals. On the contrary, there are many examples of psychology results published in *Science* and *Nature* proving unreplicable; the money-priming saga, for instance, would probably

never have happened if the original report had not been published in *Science*, and other high-profile instances of unreplicable results relating to supposedly unconscious mental processes are easy to find.[29]

More worrying, there is even emerging evidence that some aspects of quality are inversely related to journal impact factor. The Center for Open Science has recently produced a ranking of science journals including many psychology ones according to the efforts they are making to promote transparency by requiring all articles to provide open data and materials, by supporting or even requiring preregistration, and so on. Against a maximum score of 30, both *Nature* and *Science* score a distinctly moderate 11. Another ranking system for journal quality, the *N*-pact factor,[30] ranks journals by one of the key factors we discussed earlier in this chapter: the average sample size of each experiment. The rationale is that everything else held equal, studies with higher statistical power are better ones, and hence a journal publishing such studies is fostering high-quality research. Evidence again suggests a minimal correspondence between journal impact factor and this quality index.[31]

Journal editors could easily and rapidly change the prevalent culture by requiring authors to adequately power, preregister, and replicate their experiments; make all their data and code openly available, and so on. But they fear that authors would go to competitors and their journal would lose market share and prestige. Editors are no different from other scientists in reacting to prevailing incentives. A particularly stark illustration is that some editors (thankfully a very small proportion) coerce authors, prior to accepting their submitted manuscripts for publication, to include in the bibliographies citations to articles previously published in that journal, for no other reason than to boost the journal's impact factor. Many journals and their editors (a rather larger proportion) are also immensely reluctant to publish corrections or retractions of articles shown to be faulty, presumably for fear of harming the journal's brand.[32]

Now fast-forward to the prevalent current publishing model, and we see that things may be getting even worse. In the new model—for excellent reasons to do with openness—publishers receive income not from subscriptions but from per article fees. Under this "open access" model, when a journal agrees to publish an article, the researcher or her university or grant funding agency pays a processing fee to the publisher, often over $1,000 (for the journal *Nature*, the eye-watering fee is over $10,000). On publication, the article

is then publicly accessible by anyone, so the model achieves the laudable aim of making research universally available. The publisher justifies the fee by reference to the costs associated with editorial and production work, artwork, maintaining the digital repository, the cost of marketing the journal, and so on. Unfortunately, this model opens up an easy opportunity for unscrupulous businesses to launch "predatory" journals, which exist solely to make a profit and have no interest in academic rigor. Such journals bombard researchers with emails enticing them to submit their work, but in reality they carry out virtually no gatekeeping role in regard to standards, as is made abundantly clear by the many examples of such journals agreeing to publish hoax articles. One journal, for instance, published a deliberate hoax purporting to demonstrate that eating chocolate is a way of losing weight.[33]

The scientific ecosystem brings many forces to bear that support and promote poor-quality research. One effect of this has been to grossly distort our understanding of consciousness, although the consequences span probably the whole of science. Nevertheless, the growing recognition of these problems, the emergence of journals dedicated to fostering transparency, and the rapid increase in replications, preregistered or otherwise, over the past few years gives some reassurance that the culture in science is changing. Indeed surveys of economists, sociologists, psychologists, and political scientists confirm an emerging change toward the adoption of and support for practices designed to foster transparency, such as preregistration and making data, materials, and analysis code openly available. Whereas a minority of researchers adopted any of these practices fifteen years ago, far more do today.[34] But there remains a very long way to go.

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Newell, Benjamin R., 1972– author. | Shanks, David R.
Title: Open minded : searching for truth about the unconscious mind / Ben R. Newell and David R. Shanks.
Description: Cambridge, Massachusetts : The MIT Press, [2023] | Includes bibliographical references and index.
Identifiers: LCCN 2022038725 (print) | LCCN 2022038726 (ebook) | ISBN 9780262546195 (paperback) | ISBN 9780262375368 (epub) | ISBN 9780262375375 (pdf)
Subjects: LCSH: Subconsciousness. | Thought and thinking. | Self-consciousness (Awareness)
Classification: LCC BF315 .N479 2023 (print) | LCC BF315 (ebook) | DDC 154.2—dc23/eng/20230316
LC record available at https://lccn.loc.gov/2022038725
LC ebook record available at https://lccn.loc.gov/2022038726