This is a section of doi:10.7551/mitpress/14922.001.0001

Open Minded

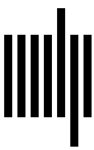
Searching for Truth about the Unconscious Mind

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

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

9 Research Biases

The achievements of science, technology, and medicine are all around us. Minor injuries or diseases that our ancestors would not have survived are routinely treated with surgical procedures, antibiotics, and other medicines. The science that makes a smartphone work is at the cutting edge of many fields: just the batteries themselves represent decades of research, culminating in the lithium-ion battery whose inventors won the 2019 Nobel Prize for chemistry. Our knowledge of the building blocks of nature derives from centuries of discoveries (such as the atom) and inventions (like the microscope).

Much of this science and technology is subject to immediate confirmation or refutation. We know that the science behind lithium-ion batteries is correct because they work. We know that ibuprofen is an analgesic because after taking it, our headache or other painful condition becomes less painful. The idea that a technology company could successfully market a new form of battery that in fact didn't work seems absurd. But despite these extraordinary successes, no one can seriously deny that substantial parts of science are in a state of crisis and are, to a large extent, broken. As soon as we move away from applications that provide immediate feedback about the veracity of the underlying science, we discover that there are no guarantees. The systematic way scientists work, collecting and analyzing data, and submitting their findings to journals for peer review, is not an infallible process for gradually accumulating correct knowledge about the world. Far from it. The scientific ecosystem, including grant-awarding panels that fund research, the research process itself, peer review, the institutional career progression and promotion mechanisms by which scientists are rewarded, and other components, is skewed in ways that lead to the generation of vast swathes of junk science. Much of this junk science concerns the unconscious.

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

This claim might seem extreme, but there are many reasons to believe it. In chapter 3, we described the phenomenon known as priming, where our perceptions or judgments or decisions can be influenced by seemingly unrelated events. Reading the word *bread* can induce a carryover effect and make us faster to subsequently read the word *butter*, and identifying a dalmatian dog in a black-and-white image (figure 3.1) can make it easier and faster for us to see the same dog in the image months or even years later. A rather more surprising form of priming was first reported and given a catchy title ("money priming") in 2006 by Kathleen Vohs and her colleagues and studied in dozens of subsequent reports (a recent review of this literature identified—incredibly—nearly 250 experiments, most of which found the effect).[1] The typical observation is the apparent modification of people's behavior on a variety of measures following exposure to images of money or tasks that involve subtle activation of the concept of money. For instance, the original study claimed that money priming causes people to work harder on difficult tasks and to become less willing to help others.[2] If this is true, the idea that workers can be nudged to exert more effort simply by subtle reminders of money is a distinctly nontrivial discovery, as is the finding that playing with coins makes children more selfish. Related research has claimed that subtle reminders of achievement, such as a photograph of a woman winning a race, can have a similar effect. In one study, for example, showing this photograph to employees in a fundraising call center increased the amount of money they raised.[3]

Later research claimed that viewing images linked to money (such as pictures of $100 bills) made people more likely to endorse free market values and social inequality.[4] They became more likely to agree with statements such as, "Some groups of people are simply inferior to others," for instance. Priming effects of this sort are explained by the unconscious activation of concepts (the mental idea of money in this case) and other closely related concepts.[5] We discuss money priming at some length in this and the next chapter for several reasons. Priming effects have been extremely influential in recent claims about the power of the unconscious mind and so deserve close scrutiny. Money priming, one of the most straightforward and intensively studied priming effects, is a veritable petri dish for considering the many biases, and the remedies for those biases, that have been identified in behavioral research over the past few years. As such, it stands as a revealing case study.

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

It seems extraordinary to imagine that something like money priming, documented time and time again in peer-reviewed journal articles, could be anything other than a true effect. Of course, there will always be limits to any phenomenon, and one would expect some money priming experiments to be failures. If the time interval between the money prime and the behavioral measure is too long, or if the prime is imperceptible or too subtle, surely the effect will become too diluted to be detectable. But this is not the problem here. Instead, the entire edifice of research on money priming is built on sand. There is (almost certainly) no genuine money priming effect.

Several lines of evidence point to this conclusion. After the initial flurry of studies on the phenomenon, researchers eventually undertook several very large efforts to replicate the early findings, and these efforts proved to be strikingly unsuccessful (we discuss these in detail in the next chapter). As doubts about this variety of priming began to accumulate, more and more negative findings made their way into journals. At the same time, questions were raised about the original study, and various methods that have been developed for identifying irregularities in bodies of research were applied to the money priming literature, indicating quite severe problems.[6] These methods are part of the set of tools, used in many scientific fields, called meta-analysis, which seeks ways of aggregating data from multiple studies. Intriguingly, an application in the field of parapsychology, the study of anomalous psychic phenomena such as telepathy, clairvoyance, and extrasensory perception, is widely recognized as the first modern meta-analysis. In the 1940s, the famous founder of parapsychology, J. B. Rhine, and his colleagues combined the results across over one hundred experiments on extrasensory perception, controversially concluding that it is a genuine phenomenon.

In combining multiple similar experiments to form an estimate of the average size of an effect, due heed needs to be paid to the possibility that the experiments that make their way into journal reports may not reflect all of the experiments that have been conducted. Suppose you conduct a test of extrasensory perception—for example, by asking a "sender" to look at a series of cards, each with one of four symbols on it, and a "receiver" to guess on the basis of the sender's transmitted thoughts which of the symbols is depicted on the card. Over a long series of trials and perhaps across many pairs of senders/receivers, you find that the receiver's accuracy is close to the level expected by chance, 25 percent. You write up your findings and send them to a prestigious journal such as *Science* or *Nature*. Your wait for a reply

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

is likely to be brief and disappointing. If you think that academic journals exist for purely scholarly purposes, then reflect on the fact that Elsevier, one of the largest journal publishers in the world, made an annual profit of over $2 billion in 2021. Journals are a competitive business, and their publishers and editors strive relentlessly to increase their profile, readership, and revenue by publishing important and attention-grabbing scientific discoveries.

Your report with its low-key findings will not exactly get the editors excited. After trying with half a dozen other journals, you may decide to give up. In so doing, you have inadvertently illustrated the *file drawer problem*. This describes a bias in which the findings that make their way into the published scientific literature are incomplete, and not just incomplete in a random way: unsuccessful experiments and studies—the ones that fail to find a difference between two groups or some other meaningful difference—are much more likely to end up in the file drawer than successful ones. The inevitable consequence of this is that the published literature presents a biased and inaccurate glimpse of the truth. If only experiments that find extrasensory perception are published, while those (perhaps vastly more) that fail to find it are left hidden from view, we will end up believing something that's not true.

But if the failed experiments are languishing out of view in scientists' file drawers, how can we ever know that they exist? Counterintuitively, by examining studies that do get published, we see traces that are highly suggestive of the existence of unpublished ones. Rhine and his associates were among the first to devise methods for dealing with the file drawer problem, but since then, numerous more sophisticated tests have been developed, and when they are applied to money priming, they provide strong grounds for believing that around the world, many researchers' file drawers must contain failures to detect the expected priming effects.

The Telltale Signs of Publication Bias

Figure 9.1 illustrates one such test. Each black circle in the figure relates to one published money priming experiment.[7] The left-hand axis represents the precision or standard error of the experiment, a statistical concept that depends on how big the experiment's sample size is (the number of participants in the experiment). The axis is presented in reverse, such that studies with higher precision (because they used larger samples) appear toward the

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

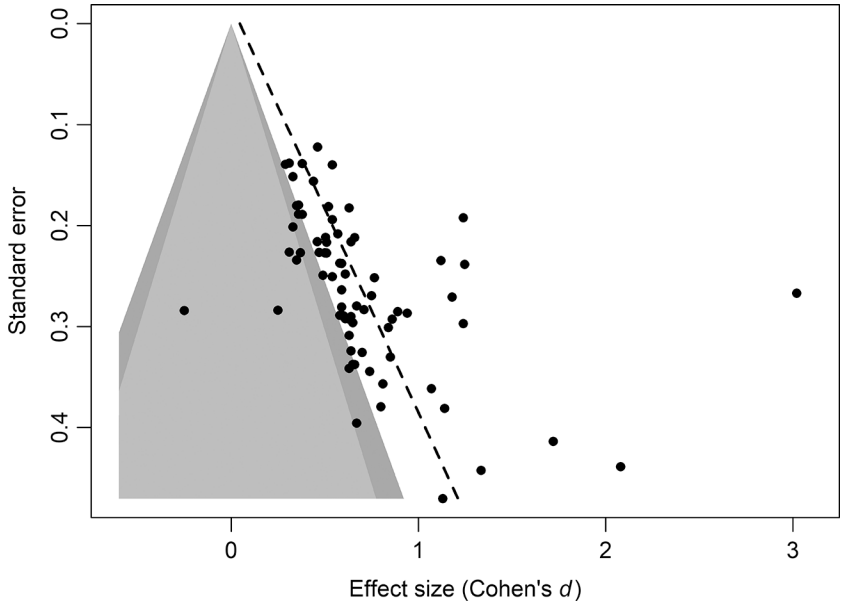


Figure 9.1
A funnel plot of data from a large number of money-priming experiments, each indicated by a dot. The vertical axis represents a measure that is related to the experiment's sample size (experiments higher up on the axis have larger samples). The horizontal axis represents the size of the money-priming effect (usually compared to a control group not shown the money prime) in a standardized measure, Cohen's *d*. Experiments finding a bigger priming effect fall further to the right. The figure clearly shows a relationship between these two measures: as an experiment's sample size gets smaller, its effect size increases. All 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 suggests that if an experiment is run with a very large sample, it will yield a priming effect close to 0 (apex of the funnel).

top of the graph. For example, the highest black circle on the graph comes from an experiment with a total sample of 275 participants, a reasonably large study, while the lowest point comes from an experiment with a mere 21 participants. The horizontal axis reflects the size of the effect obtained in each study. This is a common standardized measure known as Cohen's *d*, after the statistician Jacob Cohen, who pioneered many of our current approaches to scientific inference. For reference, think of a very obvious effect of one

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

factor on a given measure, such as the effect of sex (male/female) on a person's height. The size of this difference measured by *d* is a little less than 2 in the global population. There is, of course, considerable variation in men's height and equally in women's height, but on average, men are taller than women. Cohen's *d* quantifies this difference relative to the variation among men and women. Most of the effects shown in the figure are rather smaller than this, as one would expect from behavioral research. An effect size of 0 reflects no difference between groups or no influence of a factor on the measure of interest.

It is clear that the majority of money-priming studies (in fact, all but one of the seventy-five experiments included in figure 9.1) yield a positive effect, meaning that they find that subtle suggestions of money make people work harder on boring or difficult tasks—or make them more selfish. If we were to average the effects sizes, it is clear that the resulting aggregate or "meta-analytic" effect size would be appreciably larger than 0. But this is not the key pattern in the figure; instead, it is the evident relationship between effect size (horizontal axis) and sample size (vertical axis). Studies with smaller samples tend to obtain larger effects, and the points tend to fall either toward the top left or lower right of this inverted funnel plot. This is not the pattern that we would expect. Larger studies should yield a more precise effect size estimate than smaller ones, but the points in the funnel should be distributed symmetrically. Think of this in terms of estimates of the average height difference between men and women. Very large studies should always yield estimates quite close to the true value (just under 2 in Cohen's *d* units). They may differ slightly due to random factors (the sample may by chance contain too many unusually tall women or too many unusually short men), but the large samples mean that such randomness should be averaged out. In contrast, small studies, including only one or two dozen individuals, for example, will inevitably yield noisy estimates of the true population difference, sometimes considerably overestimating and sometimes underestimating it, but the frequency of over- and underestimations should be about equal, giving rise to a symmetrical funnel shape.

What then explains the missing points in the figure, from experiments in which small samples yielded small effects (the lower left of the figure)? One answer is the file drawer effect: such studies exist but are languishing unpublished in researchers' filing cabinets. They are languishing there because they failed to yield a statistically significant effect, the famous *p*-value whereby a

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

difference is only deemed to be "real" if (roughly speaking) the likelihood of obtaining a difference of that magnitude or greater by chance is lower than 1 in 20 ($p = .05$). The gray funnel area in the figure represents all combinations of effect size and sample size that yield statistically nonsignificant ($p > .05$) results (the dark gray region indicates marginally significant effects falling between $p = .05$ and $p = .10$). It is remarkable that the gray area so neatly separates a blank area where very few published findings exist from a cluster of published findings. In a nutshell, by examining published research, we can see a telltale pattern (asymmetry in the funnel plot) that is highly suggestive about the existence of unpublished research. This pattern is revealed only when we look at a large set of studies; it can't be seen in the individual studies themselves. This is a powerful demonstration of the importance of meta-analysis in research evaluation.

Of course in many situations we won't know anything in detail about these unpublished experiments, short of putting out a public call for scientists who work in the field to respond with information about any unpublished experiments they've undertaken on a given topic. Occasionally such calls are circulated, and indeed this has been done with respect to money priming.[8] What would we anticipate regarding these unpublished studies? Obviously the main expectation is that many of them were not published because they failed to obtain any sign of a money priming effect. (Others perhaps were unpublished for perfectly good but unforeseen reasons such as the researcher was unable to complete the experiment as planned or employed an outcome measure that proved to be unreliable.)

Is that what we see when we examine these unpublished experiments? Indeed it is. In another analysis of money-priming research, only about 35 percent of unpublished money-priming experiments obtain statistically significant results, compared to about 63 percent of published ones, and, moreover, unpublished ones yield an average effect size that is much smaller (about one-third the size) compared to published experiments.[9] A particularly striking confirmation of this relates to a form of priming that is a close cousin of money priming. In flag priming, a brief view of the American flag (it is claimed) unconsciously nudges individuals to be more right wing in their reported attitudes and voting intentions, even across a very long delay of eight months.

This quite eye-catching phenomenon was first described in a pair of experiments reported in 2011 by Travis Carter and his colleagues.[10] In a

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

rather remarkable turn of events, Carter and his collaborators later (in 2020) opened up their file drawer to provide a frank peek into the publication habits of a single research team. In the years after their initial report, they conducted many more flag-priming experiments but published none of them. As they acknowledge, this was probably due to "motivated" reasoning—namely, finding reasons for not publishing the studies, reasons that happened to align with their motivation to believe that this form of priming is genuine. It is very easy for any researcher to tell themselves that an unsuccessful replication—perhaps conducted by a student new to the team and to experimental research—should be discounted because of poor execution or some other problem. Under this harshly revealing spotlight, the contents of Carter's file drawer make it clear how pernicious this bias can be: while their two published experiments obtained an average effect size of about $d = 0.33$ (by the standards of behavioral science research, a meaningful effect), only one out of thirty-three unpublished experiments obtained a statistically significant effect and the average size of these effects was virtually zero. The fact that only the successful experiments were published means that the true status of flag priming was impossible to determine. When all experiments are combined, there is no priming effect in the totality of experiments conducted by Carter and colleagues, and other replications point to the same conclusion.[11]

This finding about money and flag priming is far from atypical. In a survey of over eighty meta-analyses in education and psychology that included both published studies and relevant unpublished ones solicited by extensive searches and well-broadcast appeals, the effect size calculated across the unpublished research was markedly smaller than that calculated across published research.[12] The conclusion is clear: if we focus only on research that makes its way past peer reviewers and editors and into journals and are not able to scrutinize unpublished research, we will be looking at a biased and unrepresentative snapshot of the truth: studies that have been cherry-picked on the basis of obtaining positive effects. The peer-review process has many virtues and helps to weed out poor-quality research, but it also introduces an unintended bias: published research will often present an overly optimistic picture of the evidence.

Indeed the cherry-picking in the case of money priming is so extreme that it miraculously turns a noneffect into an effect. Like flag priming, there is almost certainly no money-priming effect in the conditions that prevail

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

in these experiments. If we fit a trend line to the data points in figure 9.1, we can ask what the expected money-priming effect would be in an experiment with a very large sample, that is, one with a standard error near 0. The dotted line in the figure clearly suggests that this effect size would be very close to 0—the line almost touches the apex of the gray triangle, indicating a Cohen's *d* of 0. This seems like a bit of magic: from a set of experiments, almost all of which find a positive money-priming effect, we can extrapolate to what the effect would be in an "ideal" experiment, and determine that this effect would be negligible. In the next chapter, we will see that preregistered experiments designed from the outset to eliminate any possibility of bias confirm that money priming is not a genuine effect.

Money and flag priming are just two examples of a class of effects that have been labeled "social" or "behavior" priming. Other varieties, also with catchy names, include "intelligence" priming (in which individuals answer more general knowledge questions correctly when they previously thought about what it would be like to be a professor), "romantic" priming (images or text about romantic situations make men more willing to take risks), "religious" priming (subtle activation of the concept of God renders people more willing to behave prosocially), and many others (you get the idea). These represent controlled laboratory experiments that model everyday situations in which subtle cues or events might nudge our behavior unconsciously, like the claim that in-store aromas motivate us to spend more money. Like money priming, these other effects have not withstood closer scrutiny and are probably nonexistent.[13] The purported demonstrations of these effects likely represent spurious findings contaminated by publication bias and the creative employment of researcher degrees of freedom.

Researcher Degrees of Freedom

The existence of unpublished experiments obtaining smaller effect sizes than published studies is not the only possible explanation for the asymmetry seen in the funnel plot. Another possibility is that researchers might engage, perhaps inadvertently, in practices that exploit so-called researcher degrees of freedom, another avenue for bias to enter the research process.[14] Consider the following seemingly innocuous scenario. A researcher is interested in whether there is a difference in behavior between two groups, perhaps in a money-priming experiment. For one group, subtle reminders of money

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

are shown, whereas for the control group, they are not. The researcher measures (perhaps via a questionnaire) how willing participants assigned to the two groups are to engage in some volunteering activity. Suppose that there is no true priming effect. Although most of the time the experiment will correctly find no difference between the groups, occasionally—purely as a result of random fluctuations in the data—it will spuriously find a priming effect. Initially the researcher recruits twenty participants for each group and then examines the data, finding that her hypothesis seems to be confirmed: willingness to volunteer is lower in the group primed with reminders of money. However, this effect is quite small and doesn't reach the conventional threshold for statistical significance. The researcher is confident that her observed effect is genuine and that topping up her groups will reach the statistical significance threshold. She therefore tests another twenty participants in each group, reanalyzes the resulting data (now with forty participants per group), and finds a difference in willingness-to-volunteer scores that now meets the $p < .05$ threshold. She writes up her results for a prestigious journal.

The problem is that, innocently, the researcher has exploited a researcher degree of freedom (in this case, deciding how many participants to test in each group, the sample size, based on the results) in such a way as to bias her findings. Suppose the difference hadn't reached statistical significance after forty participants per group; she would probably have tested yet more, and so on until exhausting either her pool of participants or her patience. But clearly this "optional stopping" procedure inflates the probability that a purely random difference between the groups will emerge and be mistaken for a true difference. Indeed if carried on indefinitely, this procedure of repeatedly topping up and peeking at the data is guaranteed to yield a statistically significant difference, even when none exists in the population, because the experimenter will inevitably encounter one of the random fluctuations and end up capitalizing on chance rather than detecting a genuine effect. If you throw a dart once at a small target, the chances of hitting it are low. But if you throw the dart one hundred times, you would be very unlucky not to hit the target eventually.

Or consider another way in which flexibility in carrying out an experiment can lead to spurious findings. Imagine that another researcher measures how hard participants are willing to work on a boring task. After testing twenty participants in the money-primed and control groups, he observes a

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

small but statistically nonsignificant difference in the number of minutes participants in each group are willing to work on average. Noting that the difference is in the expected direction, he looks carefully at the individual data and sees that one participant in the control group works for an unusually long time while one in the primed group works for an unusually short time (the hypothesis being that money primes people to work harder). He therefore treats these data points as outliers (following perfectly sound statistical practice for excluding rogue data) and drops them from his analysis. Now the group difference reaches the magical $p < .05$ threshold, and he writes up his results for a prestigious journal.

This researcher has also exploited a researcher degree of freedom—in this case, the precise rule for treating observations as outliers. With many different choices that can be made regarding the precise outlier rule, as well as other similar decisions about transforming the data (again good statistical practice), he is boosting the probability that a random difference will look like a meaningful, nonrandom one. Some rather evocative terms are often used to refer to the many choices researchers can make that can increase the likelihood of obtaining an apparent effect in their data, even if no such effect exists, as a result of decisions taken after observing the results. One is "*p*-hacking," meaning the various tricks that a researcher can try to push a set of data over the magic $p < .05$ threshold. Another is the "garden of forking paths," from the title of a story by the Argentine writer Jorge Luis Borges, by which statistician Andrew Gelman characterizes the numerous different pathways researchers can take in analyzing their data, some of which might lead to spurious differences being obtained.

Whatever one calls these practices, they have the consequence of moving an experimental result that "should" be inside the gray funnel in figure 9.1 to a location outside the funnel.

While researcher degrees of freedom and *p*-hacking are descriptive of particular behaviors on the part of scientists, well-known concepts such as confirmation bias and motivated reasoning may be invoked to explain psychologically why these behaviors occur. When researchers set aside failed experiments and consign them to their file drawer but publish their successful experiments, they may be falling prey to confirmation bias—the tendency to search for and favor information that supports one's beliefs in preference to disconfirming information. One particular variety of this bias takes

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

the form of experimenter expectancy effects. The experimenter believes so strongly that a particular outcome will occur (and perhaps even wants that outcome to occur) that they unintentionally influence the participants in the study to conform to that expectation. You may recall the discussion in chapter 3 of research showing that experimenters who expected research participants to walk more slowly down a corridor indeed observed this outcome, and it is precisely to avoid such effects that double-blind procedures—in which both experimenters and participants are kept unaware of information that could bias their behavior—are employed in most medical trials. Expectancy effects are rife in behavioral research, including in experiments on priming.[15]

As several surveys have documented, many researchers (ourselves included) admit to having carried out practices at some stage in their careers that exploit researcher degrees of freedom. We emphasize that these practices can be and usually are entirely innocent; a researcher can in all honesty believe (and have good grounds for believing) that increasing the sample size will allow a true effect to emerge or that an observation is an outlier. Indeed our academic mentors sometimes positively encourage us to do so. We encountered Daryl Bem in the previous chapter in the context of his evidence that participants in his experiments could predict what picture was about to be displayed on a computer screen. Aside from this controversial work, Bem is a social psychologist famous for many ground-breaking contributions to research on topics like cognitive dissonance. He wrote the following in an influential guide to student researchers on how to write a journal article:[16]

> Examine them [the data] from every angle. Analyze the sexes separately. Make up new composite indexes. If a datum suggests a new hypothesis, try to find additional evidence for it elsewhere in the data. If you see dim traces of interesting patterns, try to reorganize the data to bring them into bolder relief. If there are participants you don't like, or trials, observers, or interviewers who gave you anomalous results, drop them (temporarily). Go on a fishing expedition for something—anything—interesting.
>
> No, this is not immoral. . . . In the confining context of an empirical study, there is only one strategy for discovery: exploring the data. Yes, there is a danger. Spurious findings can emerge by chance, and we need to be cautious about anything we discover in this way. In limited cases, there are statistical techniques that correct for this danger. But there are no statistical correctives for overlooking an important discovery because we were insufficiently attentive to the data. Let us err on the side of discovery.

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

This sounds strikingly like a call to undertake a fishing expedition with one's data in every conceivable way until an interesting pattern emerges. This is exactly the kind of behavior that can introduce bias into the research process and increase the likelihood of spurious findings. To be fair to Bem, he does make it clear in his guide that he is referring to *exploratory*, discovery research where novel hypotheses are being formulated and new insights sought, and he emphasizes that this is different from *confirmatory* or justificatory research, where a clear hypothesis is being put to the test and all data analysis decisions are made in advance of seeing the results, thus reducing the chances of bias. It is indeed perfectly reasonable to probe one's data in every conceivable way in the search for a brilliant new discovery or insight, provided one is transparent about doing so and the crucial pattern is replicated and confirmed in a purely confirmatory follow-up study. But he goes on to recommend that "the data may be strong enough to justify recentering your article around the new findings and subordinating or even ignoring your original hypotheses." Nothing could illustrate the crisis of scientific credibility better than this advice to present exploratory research as if it's confirmatory.[17]

What Bem is recommending is the practice known as HARKing, for hypothesizing after the results are known.[18] HARKing means reporting a hypothesis that in reality emerges from a set of data as if it were formulated before the data were collected. It is bad science because it can radically change the credibility of a pattern in a set of data. If I hypothesize in advance that a money prime will render people less willing to help others and my experiment confirms this prediction, then the hypothesis rightly gains considerable support. It would then be wholly reasonable to expend time and effort weaving the hypothesis into a larger theoretical framework. But if the hypothesis was only derived after the data were analyzed—and perhaps the data had to be massaged in complex ways before emerging—then it gains almost no support from the data. The data cannot both form the basis of the hypothesis and provide support for it. This would be circular.

In a nutshell, there are many ways in which scientists can run their experiments and analyze their data, and the ensuing garden of forking paths means that they are highly likely eventually to find something spurious in the data that looks meaningful and (more important, from the scientist's point of view) publishable, even if in reality what they've "found" doesn't exist. We contend that this is what has happened in many areas of research on the unconscious, but we emphasize that these problems probably exist

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

across the entire breadth of the sciences. There is abundant evidence from surveys that chemists, biologists, medical researchers, and those from many other disciplines admit *p*-hacking. The net result is that a high proportion of "findings" in science are likely to be misleading or even outright false.[19] Money priming provides a compelling example.

Although the evidence is striking, it is somewhat indirect. The asymmetry of the money-priming funnel plot strongly points to publication bias and the exploitation of researcher degrees of freedom, and researchers' responses to surveys make it fairly clear that questionable research practices are rife, but nonetheless these forms of evidence fall short of demonstrating concrete, irrefutable examples of poor practices such as *p*-hacking.[20] Fortunately we don't have to rely solely on these arguments, as there are now unequivocal illustrations of *p*-hacking in many specific pieces of research. One group of investigators took advantage of the fact that platforms for distributing questionnaires and collecting survey responses sometimes require all questionnaires and data to be made publicly available.[21] Hence it is possible to compare the eventual published journal article reporting a survey against the complete questionnaire that was administered. This contrast yields a stark outcome. A sizable proportion of published studies failed to mention all of the different experimental conditions in the survey. Why would this occur? The obvious reason is that a condition failed to yield the findings that the researchers expected, and they conveniently omitted it from their report. Even more startling was the finding that a majority of studies failed to report all of the measures collected in the survey, again presumably because the results were inconveniently at variance with the researchers' expectations and didn't fit into the nice story they wanted to tell. If experiment 1 yields several results that fit with the researcher's theory, but experiment 2 confirms only some of these results, then how convenient is it to pretend that the unwelcome negative findings in experiment 2 just didn't exist and that the experiment never tested these outcomes?

In addition, the results that did make their way into journal publications were much more likely to reach the $p < .05$ threshold for statistical significance than those that did not. This is rather incontrovertible evidence that researchers cherry-pick the findings that fit into the story they want to sell. When results fail to support their hypotheses, they disappear as if they were never part of the study.

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

Analyses in the Multiverse

In the standard scientific publishing model, researchers carefully describe their methods and then go on to explain the findings and their statistical interpretation of those findings, but only a single analytic method is described. The researcher chooses a single rule for dropping participants from the study, chooses a single way of dealing with outlier observations, chooses a single statistical test, and so on. Each of these choices offers scope for the exploitation of researcher degrees of freedom or *p*-hacking. Another growing trend to minimize the harm of *p*-hacking is to report the effects of making different decisions at each of these choice points, in what is called a "multiverse"(or "sensitivity") analysis. If a finding is genuine and robust, then it should still be evident even when all sorts of different choices are made about how to analyze the data. Conversely, a finding that depends critically on one specific set of choices (one route in the garden of forking paths) and disappears if any of these choices is changed is not a robust one that should be relied on for theory or practice.

Consider the following simple question: Are referees in soccer matches unconsciously more likely to give red cards to darker-skinned than to lighter-skinned players? A red card results in the ejection of the player from the game as a punishment for a major rule violation or unacceptable aggression. It has long been suspected that racial biases play a role in such decisions, but how might one go about testing this claim? Raphael Silberzahn from the University of Sussex Business School and his colleagues set out to answer this question in an unusual way by relying on a multiverse analysis.[22] First they created a data set based on information from a sports statistics company. In this data set, information on over fifteen hundred top-division players included their skin tone (judged from photos) and their interactions over the course of their careers with each of over three thousand referees (in particular, red cards given), as well as a range of details about each player's age, height, weight, and so on. One might think that on the basis of this data set, it would be fairly straightforward to determine whether darker-skinned players received more red cards than lighter-skinned ones.

But a moment's reflection suggests that there will be quite a few forking paths in this particular garden. For example, it might be the case that darker-skinned players tend to play slightly more often in defensive positions than lighter-skinned ones (or vice versa). Defenders might be slightly

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

more (or less) likely than attackers to be given red cards. These two trends, which might be very slight, could nonetheless combine to yield a pattern in which darker-skinned players falsely appear to be more prone to punishments, but the pattern would not be indicative of an influence of skin color. So any approach to analyzing this data set will inevitably throw up a range of questions that the investigator needs to address. In an ingenious approach, Silberzahn and his colleagues simply invited expert research teams from around the globe to take on the challenge of analyzing the data set according to their own particular preferred approaches, and twenty-nine agreed to take up the challenge.

What was the outcome? No two teams reached exactly the same estimate of the effect of skin tone on the likelihood of being given a red card, and although the majority (about two-thirds) of the teams concluded that there is a relationship, many (about one-third) concluded that there isn't. The teams adopted a staggering range of analytic approaches, using numerous different statistical techniques.

The salutary point of this example is that any one of the analyses could have been undertaken individually and justifiably published in the normal way in a peer-reviewed journal. A total of twenty-nine articles would have made their way into the literature, with no clear consensus about the true answer to the question. Since all the teams analyzed exactly the same data set, we can say categorically that variation in the decisions that intelligent researchers make about how to analyze their data can lead to polar opposite conclusions. If we have no transparency about these decisions and about how robust researchers' conclusions are in the face of different sets of decisions, then we cannot reasonably evaluate the outcome of any single piece of research.[23]

The convergence of the biases discussed in this chapter yields spurious conclusions about the mind. In case the evidence we've presented isn't sufficient to convince you of this, then consider one final point. Against any reasonable scientific criteria, the quantity of evidence for paranormal phenomena such as telepathy, clairvoyance, and precognition (seeing the future) is overwhelming. Hundreds of research reports have been published in peer-reviewed journals of successful demonstrations of these phenomena. In the previous chapter, we described Daryl Bem's infamous experiments apparently showing that people can know, in advance, where an erotic image was about to be displayed on a computer screen. Although

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

we described many reasons (quite apart from their implausibility) not to believe Bem's findings, the general claim that paranormal phenomena exist rests on vastly more evidence than this one set of dubious experiments. It seems highly likely that if we collected a large amount of data relating to some putative but in reality nonexistent paranormal phenomenon and subjected those data to a multiverse analysis, at least some of the analysts would wrongly conclude that the effect is genuine.

Etzel Cardeña, a psychologist at Lund University in Sweden, has summarized meta-analyses and concluded that they provide compelling support for telepathy, clairvoyance, precognition, and related phenomena.[24] Indeed the evidence from these meta-analyses is probably—by any objective standards—at least as strong as the evidence for many standard psychotherapy treatments and numerous other widely accepted results. Cardeña takes them as proving the existence of the paranormal. For anyone with a more skeptical view of such phenomena, they confirm the existence of a raft of research practices and biases that allow scientists to fool themselves and others.

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024

MIT Press Direct

© 2023 Massachusetts Institute of Technology

This work is subject to a Creative Commons CC-BY-NC-ND license.
Subject to such license, all rights are reserved.

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

Downloaded from http://direct.mit.edu/books/oa-monograph/chapter-pdf/2152365/c004900_9780262375375.pdf by guest on 19 May 2024