Chapter 1

Five Models of Science, Illustrating How Selection Shapes Methods

Paul E. Smaldino

"The good thing about science is that it's true whether or not you believe it." This oft-repeated quote, attributed to the astrophysicist and TV presenter Neil deGrasse Tyson, was seen everywhere at the March for Science, a set of gatherings held around the world on April 22, 2017. The quote has become a rallying cry for supporters of science—and of the application of scientific knowledge in daily life—against widespread science denialism. And of course, science should be defended. Carl Sagan, Tyson's predecessor as host of *Cosmos*, noted that science not only increases our knowledge of the world but also serves as a bulwark against superstition and charlatanry (Sagan 1996). However, there is a counterpoint to Tyson's claim. Plenty of science, or at least scientific results, are *not* true.

During the first decade of the twenty-first century, the biotech company Amgen attempted to confirm the results of fifty-three published oncology papers deemed "landmark" studies. Of these, they claim to have successfully replicated only six (Begley and Ellis 2012).[1] In 2015, a team of 270 researchers calling themselves the Open Science Collaboration repeated one hundred studies from published psychology papers. Of these, they successfully replicated only thirty-nine results (Open Science Collaboration 2015). In 2016, neuroscientists discovered design errors in the most popular statistical packages used to analyze fMRI data, indicating that as many as 70% of the results obtained using these packages may be false positives (Eklund, Nichols, and Knutsson 2016). And in 2018, a team of social scientists targeted twenty high-profile studies published in the prestigious journals *Science* and *Nature* and successfully replicated only twelve; even among these, most of the effects turned out

to be smaller than originally published (Camerer et al. 2018). Indeed, a survey conducted by *Nature* in 2016 revealed that a large proportion of empirical scientists, hailing from fields as diverse as chemistry, biology, physics, earth sciences, and medicine, had failed to replicate other researchers' results (Baker 2016).

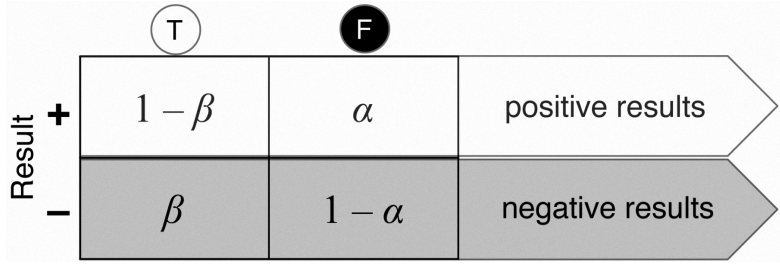
This is a problem. Our understanding of the world relies on facts. Charles Darwin understood the perniciousness of false facts, writing in *The Descent of Man*, "False facts are highly injurious to the progress of science, for they often endure long; but false views, if supported by some evidence, do little harm, for every one takes a salutary pleasure in proving their falseness; and when this is done, one path towards error is closed and the road to truth is often at the same time opened" (1871, 385). What he is saying in his overwrought Victorian prose is that we shouldn't worry too much about false theories, because academics are competitive and love to take each other down a peg by demonstrating logical inconsistencies in one another's theories. Since logic is a common language in science, the competition for theoretical explanations remains relatively healthy. However, any coherent explanation must rely on a firm foundation of facts. If our facts are false, we end up wasting our time arguing about how best to explain something that isn't even true.

Science involves both theory building and fact finding. This chapter focuses on the fact-finding aspect, and as a shorthand the search for facts is what I will mean henceforth by the term "science." In this sense, science can be viewed as a process of signal detection for facts. We wish to discover true associations between variables. However, our methods for measurement are imprecise. We sometimes mistake noise for signal and vice versa.

How we conceptualize the scientific enterprise shapes how we go about the business of conducting research, as well as how we strive to improve scientific practices. In this chapter, I'll present several models of science. I'll begin by showing ways in which the classic "hypothesis testing" model of science is misleading and leads to flawed inferences. As a remedy, I'll discuss models that treat science as a population process, with important dynamics at the group level that trickle down to the individual practitioners. Science that is robust and reproducible depends on understanding these dynamics so that institutional programs for improvement can specifically target them.

A First Model of Science: Hypothesis Testing

Early in our schooling, many of us are taught a simple and somewhat naive model of science as "hypothesis testing" (figure 1.1). The scientist comes up with a hypothesis about some natural system. She cannot



1.1. A first model of science. Hypotheses are investigated and results, with characteristic error rates, are recorded. The real epistemic state of each hypothesis, true or false (T or F), is unknowable except through this sort of investigation.

directly infer the essential epistemic state of the hypothesis, whether it is true or false. Instead, she investigates the hypothesis by experimentation or other empirical means, which results in either a positive result in support of the hypothesis or a negative result indicating a lack of support. The alignment between her results and the epistemic state of the hypothesis is necessarily imprecise. There is some risk of a false positive, $\alpha = \Pr(+|F)$, as well as a false negative, $\beta = \Pr(-|T)$. These outcomes are sometimes called Type 1 and Type 2 errors, respectively.[2] This uncertainty forces us to ask: How confident should our scientist be in her results?

Consider the following scenario. Dr. Pants investigates one of her many hypotheses. Using her well-tested method, the probability that the test will yield a false positive result is 5%. That is, $\Pr(+|F) = 0.05$. If the hypothesis is true, the probability that the test will correctly yield a positive result is 50%. That is, $\Pr(+|T) = 0.5$. The test is conducted, and the result is positive! Now, what is the probability that Dr. Pants's hypothesis is correct?

You may be tempted to answer 95%. After all, the probability of a false positive is 5%, and it's clear that 100 – 5 = 95. If this is your answer, you are not alone. When a version of this question was posed to students with scientific training, 95% was indeed the most common answer, at least in years past (Gigerenzer and Hoffrage 1995). Why is this wrong? Recall that we are looking for the probability that the hypothesis is true conditional on obtaining a positive result, $\Pr(T|+)$. Fortunately, we have a handy mathematical tool for computing exactly this sort of conditional probability. Using Bayes' Theorem, we can write out our conditional probability as follows:

$$\Pr(T|+) = \frac{\Pr(+|T)\Pr(T)}{\Pr(+|T)\Pr(T) + \Pr(+|F)(1 - \Pr(T))}$$

You'll notice right away that there's a term in this equation I haven't provided: Pr(T). This is the prior probability that *any* hypothesis being tested by Dr. Pants is true, often called the *base rate*. We ignore the base rate at our peril.

A Second Model of Science: Hypothesis Selection and Investigation

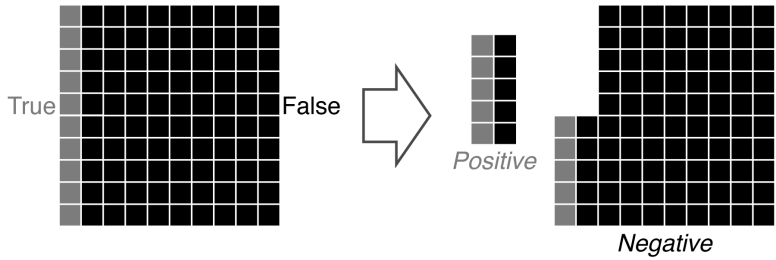
Imagine now that Dr. Pants tests not one but one hundred hypotheses. Of these, ten are true and ninety are false. If you want a more concrete example, imagine Dr. Pants runs a behavioral genetics lab. She is looking for single nucleotide polymorphisms (SNPs) that correlate with a heritable behavioral disorder. She tests one hundred SNPs, of which ten are actually associated with the disorder. Thus, the base rate is $b = 0.1$. If this seems low, consider that for many disciplines, the base rate may actually be much lower. Every association tested, every statistical test run, is a hypothesis that may be supported. Dr. Pants tests her hypotheses using the method described in the previous paragraph, with $\alpha = 0.05$ and $\beta = 0.5$. So what is the probability that a hypothesis with a positive result actually reflects a true hypothesis? In this case, it's 50%, not 95% (figure 1.2). And the lower the base rate, the lower this posterior probability gets. Worse yet, in reality we can never know for certain the epistemic states of our hypotheses, nor can we easily estimate the base rate. Our results are all we have.

So now we have a second model of science that includes the process of hypothesis selection as well as the experimental investigation of that hypothesis (figure 1.3).[3] We can capture this model in terms of the posterior probability that a positive result indicates a true hypothesis using the notation introduced so far:

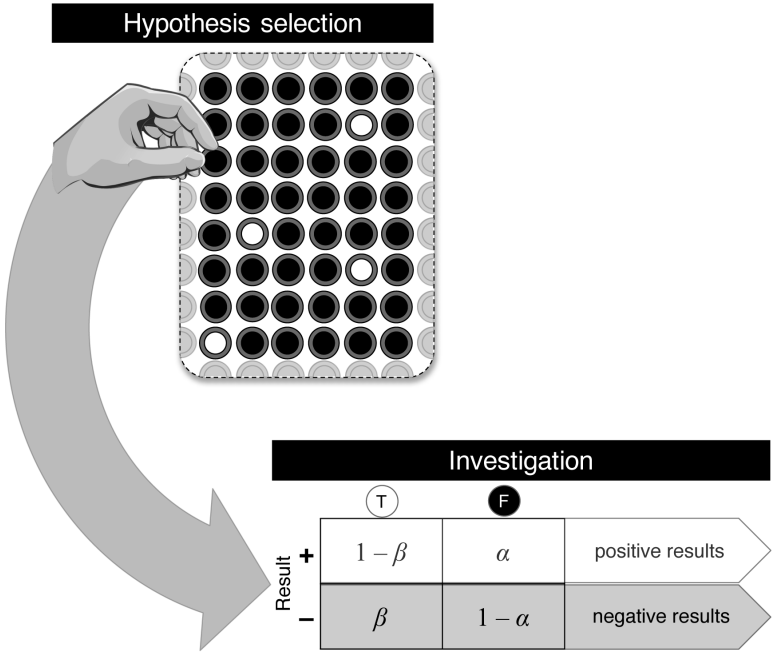
$$\Pr(T|+) = \frac{(1-\beta)b}{(1-\beta)b + \alpha(1-b)}$$

This Bayesian model of science was introduced by Ioannidis (2005) in his now classic paper, "Why Most Published Research Findings Are False." The analysis is straightforward. If the base rate, b, is low, then even a moderate false positive rate (such as 5%) will lead to a low posterior probability and a large number of false positives.

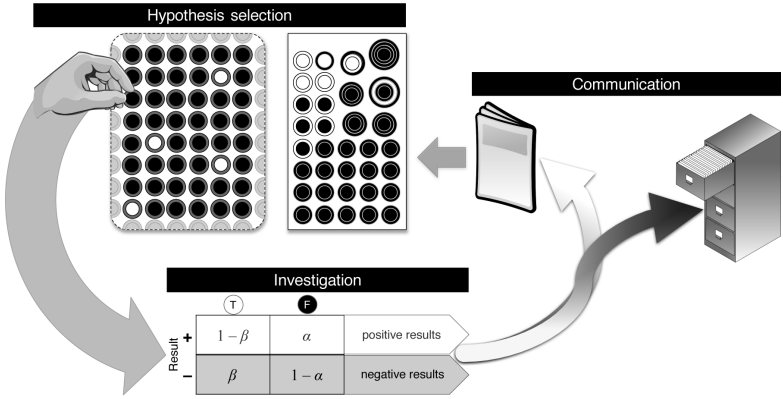
One concern about this model is that it treats each hypothesis in isolation. It ignores the social and public aspect of science. Scientists don't just produce results; they also try to publish them, and some results are easier to publish than others. Once published, results can then be replicated, and with new information comes the opportunity for new estimates of the epistemic states of the underlying hypothesis.



1.2. The importance of base rate. Left: 100 hypotheses are tested, of which 10 are true (the base rate is $b = 0.1$). Right: 50% of the true hypotheses and 5% of the false hypotheses yield positive results, producing a posterior probability that a positive result is actually true of approximately $\Pr(T|+) = 0.5$.



1.3. A second model of science. Investigation is preceded by hypothesis selection. The inner circles indicate the real epistemic value of each hypothesis. Black indicates false, white indicates true. The gray outer circle represents the fact that these epistemic values are unknown before investigation.



1.4. A third model of science. After hypothesis selection and investigation, results are communicated. Some results end up published and become part of the literature, which can accrue through replication. This is indicated by the rectangle containing concentric circles. Each layer represents a positive (white) or negative (black) results. Results that are not published end up in file drawers, unknown to the scientific community.

A Third Model of Science: The Population Dynamics of Hypotheses

The first two models of science both portray a science in which each hypothesis is investigated in isolation. But consider what happens to a result once the hypothesis has been investigated. The researcher will sometimes decide to publish the result. I say "sometimes" because some results are never published, especially when they don't support the hypotheses being tested. These results end up in the "file drawer" (Rosenthal 1979). Once published, the studies supporting a given hypothesis can be replicated, whether by other labs or by the one that generated the original result.

Our third model conceptualizes hypothesis testing as a dynamical system involving a large number of hypotheses being tested by a large number of scientists (figure 1.4). A scientist first selects a hypothesis to test. A novel hypothesis is true with probability b, the base rate. The hypothesis is investigated, producing results. These results can then be disseminated to the scientific community via publication. This stage is important, because not all results are published with equal probability. Novel positive results are usually the easiest to publish. Negative results are published at much lower rates (Fanelli 2012), possibly due to being rejected by journal editors but also because they are viewed as carrying low prestige for researchers and are therefore rarely submitted (Franco,

Malhotra, and Simonovits 2014). Once findings are published, they can be replicated. The results can then be added to the literature, but only if they are published. As results accrue, each hypothesis is associated with a record of positive and/or negative results in the published literature. Because some types of results are more likely than others to be published, the published literature likely reflects a biased record of investigation.

This dynamical model was introduced and analyzed in an earlier paper (McElreath and Smaldino 2015). Our analysis focused on the probability that a hypothesis was true, conditional on its publication record.[4] For simplicity, we operationalized the publication record as a tally of the net positive findings—that is, the number of positive results minus the number of negative results in the published literature. Although this conditional probability was influenced to some degree by all of the model's parameters, we found that the two parameters exerting the largest influence—by far—were the base rate, b, and the false positive rate, α. If the base rate is high (so that most tested hypotheses are true) and the false positive rate is low (so that most positive results reflect true hypotheses), then a single positive result likely reflects a true hypothesis. However, as base rate decreases and false positive rate increases—to values that, I must add, I view as quite realistic for many disciplines—then more and more successful replications are necessary to instill the same amount of confidence in the truth of a hypothesis.

Above all, this indicates that replication is important for a healthy science (Smaldino 2015). Indeed, our analysis showed that replication studies are valuable even when they use designs with different methodological power than the original investigations. More than that, we shouldn't be surprised that some results fail to replicate. Some erroneous results are inevitable. When methods are imperfect, both false positives and false negatives may be common. That said, the model also illustrates that improvements to the practices and culture of science should focus on factors that increase the base rate of true hypotheses and lower the rate of false positives results, so as to decrease the number of false facts in the published literature.

A number of factors lead to false discovery. False facts are more common when:

- Studies are underpowered, because small sample sizes tend to lead to false positives and ambiguous results.
- Negative results aren't published, distorting the publication record by eliminating disconfirmatory evidence.
- Statistical techniques are misunderstood, leading to false positives and ambiguous results.

> Surprising results are the easiest to publish, because such results have a low base rate of being true, given priors to the contrary.

Although the factors in this list may be new to some readers, scientists have, in general, been aware of these issues for decades. Why, then, isn't science better? Understanding how scientific practice—and not just scientific knowledge—changes over time requires a new model that includes the scientists themselves in the model dynamics. Before introducing such a model, I'll need to say a few words about some of the incentives that structure human social behavior.

A Brief Interlude on Incentives

Science is the search for truth about the natural world, for a better understanding of our universe. Scientists, however, are also human beings who need steady employment and the resources to conduct their research. Obtaining those jobs and securing that funding is far from trivial these days. There are currently far more PhDs looking for employment in academia than there are permanent positions for them to fill. In several disciplines, including biomedicine and anthropology, the creation of new PhDs outpaces the creation of new faculty positions by a factor of five (Ghaffarzadegan et al. 2015; Speakman et al. 2018). More generally, the number of open faculty positions in scientific disciplines is only a small fraction of the number of total PhDs awarded each year (Cyranoski et al. 2011; Schillebeeckx, Maricque, and Lewis 2013). This creates a bottleneck at which selection is nonrandom. In academic science, this selection pressure is often linked to an individual's publication history, as evinced by the clichéd admonition to "publish or perish."

Successful scientists are certainly publishing more. Since just the early 2000s, the number of publications at the time of hiring for new faculty has more than doubled in fields such as evolutionary biology (Brischoux and Angelier 2015) and cognitive psychology (Pennycook and Thompson 2018). A large study of over twenty-five thousand biomedical scientists showed that scientists who ended up as principal investigators (PIs) consistently published more papers and placed them in higher-impact journals that those researchers who ended up leaving academia (van Dijk et al. 2014).

It may not be immediately obvious that preferential reward for productivity and impact factor are bad things. Indeed, it seems that we should *want* scientists to be productive and we should *want* their work to have a wide impact. Don't we want our scientists to be awesome? The difficulty is that awesomeness is in reality quite complicated and multidimensional. The importance of research may not be manifest for

quite some time, and a lack of productivity can just as easily reflect careful study of a difficult problem as it can a lack of drive. This difficulty becomes a serious problem when awesomeness is assessed with crude, quantitative metrics like paper count, journal impact factor, and h-indices. It has been widely noted by savvy social scientists that, as Campbell (1976, 49) noted, "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." When incentives to publish drive scientists, science itself may become distorted.

There is evidence that scientists do, in fact, respond to incentives. In China, as in several other countries, PIs are often given cash rewards for publishing in top English-language journals. This system began in the early 1990s with small rewards, but the size of the rewards has grown tremendously. As of 2016, Chinese researchers were paid, on average, $984 for a paper in *PLOS ONE*, $3,513 for a paper in the *Proceedings of the National Academy of Sciences*, and a whopping $43,783 for a first-author paper in *Science* or *Nature* (Quan, Chen, and Shu 2017). Correspondingly, between 2000 and 2009, Chinese submissions to the journal *Science* nearly quadrupled (Franzoni, Scellato, and Stephan 2011). China was recently declared the world's largest producer of scientific papers (Tollefson 2018). Such cash-for-papers incentives can be found in several other countries, including India, Korea, Malaysia, Turkey, Venezuela, and Chile (Quan, Chen, and Shu 2017). The West is not immune either. For example, I recently had dinner with some American psychologists, who told me with pride about how much their graduate students published. Their program provided a cash prize of several hundred dollars for the best student paper each year. When I asked how they assessed the best paper, they told me that a first-author publication in a top journal was the best indicator. "Do you read all the papers?" I asked. The answer was no; the journal's reputation was deemed a sufficient mark of quality. It is not hard to see how students in this program are incentivized not only to produce papers but to produce a particular type of paper.

Evidence that scientists respond to incentives can be more subtle. Vinkers, Tijdink, and Otte (2015) looked at relative word frequencies in PubMed abstracts between 1974 and 2014. They found dramatic increases in the frequencies of positive, congratulatory words. Frequencies of the words "innovative" and "groundbreaking" had each increased 2500%. Frequency of "novel" had increased 4000%. And frequency of "unprecedented" had increased 5000%. There are, of course, two possible explanations for this shift in word frequencies. The first is that

contemporary scientific research is actually twenty-five times more innovative than it was forty years ago. The other, a smidge more likely, is that scientists are responding to incentives to distinguish their word as important and pathbreaking.

A system that rewards novel, innovative results can—and does—incentivize cheating. Recent examples include Jan Schön in physics, Diederik Stapel in psychology, and Brian Wansink in nutrition science. A personal favorite is a case of fraud uncovered by the editors of the *British Journal of Clinical Pharmacology*. The authors of a paper claiming impressive results suggested as reviewers several prominent scholars in their field. These scholars were contacted as reviewers, and all returned glowing reviews within just a few days. One of the editors grew suspicious at the quick responses from the busy big-shot scientists, and contacted them at the email addresses listed on their university web pages. They were all surprised by the emails, because none of them had heard of the paper in question. The explanation: when the authors submitted their manuscript, they had provided fake email addresses for their suggested reviewers and submitted forged reviews of their own paper (Cohen et al. 2016).[5]

Fraud surely happens, but it's also probably the exception rather than the rule. Most scientists are well-meaning people who want to learn about the world. The problem is that incentives for maximizing simple quantitative metrics, which act as proxies for more meaningful but multifaceted concepts like productivity and influence, can be detrimental even if all actors are well intentioned. To help explain why, we'll turn to a new model of science that includes the scientists as well as the hypotheses.

A Fourth Model of Science: Variation, Heritability, and Selection

Science is a cultural process that, like many cultural processes, evolves through a Darwinian process (Richerson and Boyd 2005; Mesoudi 2011; Smaldino 2014; Smaldino and McElreath 2016). Philosophers of science including Campbell (1965), Popper (1979), and Hull (1988) have discussed how scientific theories evolve by variation and selective retention. But scientific methods can also evolve. Darwinian evolution requires three conditions to occur:

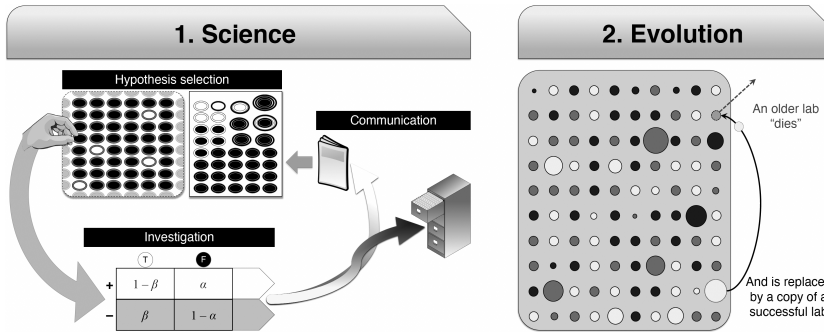
1. There must be variation.
2. That variation must have consequences for survival or reproduction.
3. Variation must be heritable.

Research practices and methods certainly vary. That variation leads to differences in the sorts of results that are produced and, consequently, the publications that arise from those results. These publications have consequences in determining who is successful in terms of getting hired and promoted, securing grants, attracting graduate students and postdocs, and placing those trainees in positions heading their own research groups. And variation in practice is partly heritable, in the sense that trainees acquire research habits and statistical procedures from mentors and peers. Researchers also acquire research practices from successful role models in their fields, even if they do not personally know them. Therefore, when researchers are rewarded primarily for publishing, habits that promote publication are likely to be passed on.

If we want to understand how we might minimize false discoveries, we need a model of science that includes variation among scientists. This model has two phases: Science and Evolution (figure 1.5). In the Science phase, each research lab chooses and investigates hypotheses and tries to publish their results, just as in our third model of science. However, the methods used by each lab can differ, which affects the rate at which they conduct research and the probability of certain results. More specifically, consider a population of labs, all conducting research. We make the following assumptions:

Each lab has characteristic methodological power, $\Pr(+|T)$.
Increasing power also increases false positives, unless effort is exerted. This is because it is easy to have perfect power if every result is positive, but correctly eliminating the false hypotheses requires additional work.[6]
Additional effort also increases the time between results because each study requires more work.
Negative results are harder to publish than positive results.
Labs that publish more are more likely to have their methods "reproduced" in new labs.

This model was first presented and analyzed in another paper with Richard McElreath (Smaldino and McElreath 2016). First, we found that if effort is held constant and power is allowed to evolve, power evolves to its maximum value and the false discovery rate (the proportion of published results that are incorrect) skyrockets. Everything is deemed "true," and we have no information about anything. This scenario is pretty unrealistic. We have fairly good ways of assessing the power of research methods, and no one would ever allow this to happen. However, *effort* is notoriously difficult to assess. If we hold power con-



1.5. A fourth model of science. Dynamics occur in two phases: Science and Evolution. In the Evolution stage, labs compete for research positions. A lab's methods are represented by its shading, and its prestige is represented by its size. When a new position opens, it is more likely to be filled by someone using the methods of more prestigious labs.

stant and allow effort to evolve, we find that effort reliably evolves to its minimum value, and once again the false discovery rate balloons. To reiterate, this dynamic requires no cheating or strategizing on the part of our agents, only that publication is a determinant of job placement. We have referred to this dynamical process as "the natural selection of bad science" (Smaldino and McElreath 2016).

What does this mean? It means that if our model of science is at least moderately realistic, and incentives for publishing do drive selection on research methods, then we should see evidence for impediments to the improvement of scientific methods on the timescale of generations. If, on the other hand, incentives are rewarding methodological rigor, we should see a steady increase in the quality of methods for scientific inquiry.

In 1967, Paul Meehl cautioned about the misuse of *p*-values, warning that scientists were wrongly interpreting their meaning and consequently generating lots of false positives (Meehl 1967). In 2016, the American Statistical Association published its "Statement on *p*-Values," cautioning about their misuse and warning that scientists were wrongly interpreting their meaning and consequently generating lots of false positives. The ASA bemoaned, "Let us be clear. Nothing in the ASA statement is new. Statisticians and others have been sounding the alarm about these matters for decades, to little avail" (Wasserstein and Lazar 2016, 130).

In 1962, Jacob Cohen published a meta-analysis of abnormal and social psychology experiments, noting the frustratingly low statistical power of most published research (Cohen 1962). He cautioned

that many studies were not sufficiently powered to adequately provide confirming or disconfirming evidence, leading to an excess of spurious results. In the late 1980s, two studies provided new meta-analyses investigating whether there had been any improvement to the average statistical power of psychological research (Sedlmeier and Gigerenzer 1989; Rossi 1990). They found no improvement. Recently, Richard McElreath and I updated those studies and confirmed that, on average, there was no improvement to the average statistical power in the social and behavioral sciences through 2011, with an average power to detect small effects of 0.24 (Smaldino and McElreath 2016).[7] Szucs and Ioannidis (2017) provided a focused study of ten thousand papers published in psychology, medicine, and cognitive neuroscience journals between 2011 and 2014 and similarly found very low power in all three fields.

The natural selection of bad science appears to be pernicious. I previously noted the importance of replication for assessing the true epistemic value of hypotheses. Could replication similarly help to curb the degradation of methods? One particularly interesting, if extreme, suggestion came from Rosenblatt (2016), who proposed that the authors of each published paper, or their host institutions, sign a contract committing them to pay a fine if their studies fail to replicate. Let me be clear: this is a terrible idea. As stated earlier, occasional failure to replicate is to some extent the price of doing business in scientific research. However, it is one of the more concrete suggestions for using replication to improve science. So we put it—or something like it—into the model.

Under our replication extension, all labs committed a proportion r of their investigations to replicating previously published results. We assumed that all replications were publishable regardless of the result and carried half of the prestige carried by a novel positive finding.[8] If another lab replicated a finding successfully, the lab that published it originally got a small boost in prestige. If another lab *failed* to replicate a finding successfully, the original authors suffered a tremendous loss of prestige. To be honest, we thought this extreme intervention would curb the decline in effort and the runaway false discovery rate. In hindsight, it is clear why it didn't. Although some labs did suffer a huge loss of prestige, the most successful labs were still those who cut corners and avoided being caught.

Incentive structures that push scientists to boost quantitative metrics like publication counts and impact factors can lead to the degradation of methods. This dynamic requires no fraud or ill intent on the part of individual actors, only that successful individuals transmit their methods.[9] From this, we might conclude that changing individual behavior—each of us improving our methods—is not sufficient to im-

prove scientific methods; this requires institutional change. Specifically, it requires that the selection bottlenecks of hiring and promotion are not overly focused on those metrics but can instead provide a more nuanced assessment of researcher quality that maintains high methodological integrity.

Unfortunately, institutional change is far from easy. For the most part, institutions are not *meant* to change easily. They provide a stable framework that structures social interactions and exchanges, and ensure some consistency to the operation of a society in the absence of enforcement by specific individuals (North 1990). This means that we run into trouble when our institutions are unhealthy. If we are to change the institutional incentives for publishing in academic science, we should be aware that such change will likely be slow. Is there anything else that can be done in the short run?

There are many efforts currently under way to improve the norms and institutions of academic science regarding rigor and reproducibility, often under the banner of the "Open Science" movement (Nosek, Spies, and Motyl 2012; Munafò et al. 2017). Some of these new norms include preregistration and registered reports (Nosek and Lakens 2014; Chambers 2017), preprints (Bourne et al. 2017; Smaldino 2017b), double-blind and open peer review (Mulligan, Hall, and Raphael 2013; Okike et al. 2016; Tomkins, Zhang, and Heavlin 2017), and better training in methods, statistics, and philosophy of science. At the same time, funding agencies are increasingly paying attention to what gets funded, and some have been shifting how they fund new research projects. How do these developments influence the conclusions from our fourth model of science?

A Fifth Model of Science: Follow the Money

Our fourth model of science makes several pessimistic—if realistic—assumptions about the way academic science works in our era. However, changes in just the last few years prompt us to challenge some of these. I want to focus on three specific assumptions and discuss what happens when we relax or alter them.

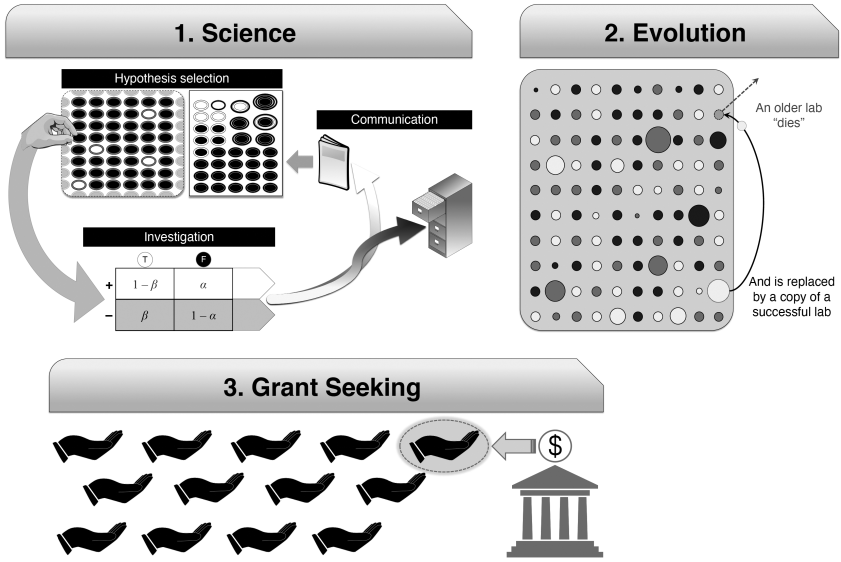
Assumption 1: Publishing negative results is difficult or confers little prestige. This assumption is realistic, because negative results are rarely published (Fanelli 2012) or even submitted (Franco, Malhotra, and Simonovits 2014). However, there is an increasingly large push to publish negative results. Many journals now accept registered reports, in which the research plan is peer reviewed before a study is conducted. Once approved, the paper's acceptance is contingent only on adherence to the submitted plan and not on the character of the results (Nosek and Lak-

ens 2014; Chambers 2017). A recent study by Allen and Mehler (2019) found that among studies using registered reports, 61% of results did not support the authors' original hypotheses, compared to estimates of 5%–20% of null findings in the wider literature.[10] What if publication bias against negative results were eliminated?

Assumption 2: Publishing positive (confirmatory) results is always possible. This assumption ignores the corrective role of peer review in maintaining high-quality research. The assumption is realistic, because there is little evidence that peer reviewers can act as effective gatekeepers against false discovery. The many failed replications discussed earlier in this chapter testify to that. Peer review may in many cases be more about maintaining group norms than about weeding out error. There is widespread evidence that peer reviewers can be biased toward prestigious individuals and institutions and against authors who are women and underrepresented minorities (Budden et al. 2008; Tomkins, Zhang, and Heavlin 2017). If peer review was reliable, we should expect consistency between reviewer recommendations. Instead, a number of studies have found low correlation between reviewer decisions on grant panels (Cole and Simon 1981; Marsh, Jayasinghe, and Bond 2008; Mutz, Bornmann, and Daniel 2012), conference proceedings (Langford and Guzdial 2015; Deveugele and Silverman 2017), and journal articles (Peters and Ceci 1982; Cicchetti 1991; Nicolai, Schmal, and Schuster 2015).

Nevertheless, we increasingly see efforts to improve the conditions that facilitate effective peer review. Registered reports remove biases based on the novelty or expectedness of a study's results (Nosek and Lakens 2014; Chambers 2017). Double-blind peer review aims to reduce biases, including those based on prestige, familiarity, gender, race, or ethnicity (Mulligan, Hall, and Raphael 2013; Okike et al. 2016; Tomkins, Zhang, and Heavlin 2017). Journals increasingly require or incentivize open data and methods, which improves the ability of peer reviewers to assess results, and the increased use of repositories such as OSF (Open Science Framework) and GitHub has helped to facilitate this behavior. Open peer review and the increased use of preprint servers also allow for a greater number of critical eyes to read and comment on a manuscript before it is published (Bourne et al. 2017; Smaldino 2017b). And better training in statistics, logic, and best research practices—as evidenced by the popularity of books, massive open online courses, podcasts, symposia, and conferences on Open Science—may promote more informed reviews. What if peer review was effective at filtering out false discovery?

Assumption 3: Research productivity is constrained only by the ability to complete projects. This assumption ignores the role of funding, which is required for much scientific research. This assumption was justified



1.6. A fifth model of science. In addition to the Science and Evolution phases, labs also compete for grant funding, which enables them to conduct more research.

by the desire to ignore differences in access to funding and focus on the bottlenecks at hiring and promotion. Moreover, if one assumes that success in securing grant funding results from success in the quantity and prestige of one's publications, then including explicit funders in the model is unnecessary. Instead, what if funders ignored publication records, or even focused on funding projects with the most rigorous methods?

The norms of hiring and promoting researchers based on simple metrics are entrenched in deeply rooted tradition and diffuse across many academic institutions; they will not be changed quickly or easily. In contrast, the recent changes highlighted above are occurring rapidly, due to greater top-down control from journals and funders. To investigate the consequences of these changes, we will once again revise our model of science.

We again consider a finite population of labs. Each lab has a characteristic methodological rigor (or lack thereof), which is linked to the false positive rate of the results they obtain. In our fourth model, a lab's productivity was limited only by its rigor. This time, investigating hypotheses requires funding. Each lab is initialized with some start-up funds it can use to conduct research. Once these funds are exhausted, additional funds must be acquired from grant agencies.

To our two phases of Science and Evolution, we add a third: Grant Seeking (figure 1.6). In the Grant Seeking phase, some of the labs apply for funding, and the one that best matches the funding agency's allocation criteria is awarded a grant. We might consider any number of strategies. My colleagues and I have considered those based on publication quantity, funding labs at random, and targeting those labs with the most rigorous methods. The Science phase looks quite similar to that of our previous models, having three phases—hypothesis selection, investigation, and communication. Here we may also take the opportunity to study changes to peer review and publication bias as discussed. In the communication phase, positive results are always published, and negative results are published with probability p. Erroneous results (in which the result does not reflect the real epistemic state of the hypothesis) are successfully blocked during peer review with probability r. The Evolution phase works exactly as it did in the previous model, such that labs with more publications are most likely to transmit their methods to the next generation. This is worth repeating: the selection pressure for publication quantity is still present. For a detailed analysis of this model, see Smaldino, Turner, and Contreras Kallens (2018). Here, I summarize our main results.

First, we can ask whether, in the absence of any contributions from funding agencies, curbing publication bias and improving peer review can promote substantial improvements to reproducible science. There is bad news, then good news, and then bad news again. The bad news is that, taken one at a time, each of these improvements must be operating at nearly maximum levels for any improvements to occur. That is, negative results must be published at equal rates as positive results, and peer reviewers must be nearly perfect in detecting false discoveries. The good news is that the effects of these two interventions are additive, so that moderate improvement to both publication bias and peer review *can* decrease the rates of false discovery to some extent. The bad news (again) is that this effect operates on the published literature, so that more published results are true, but does little to improve the quality of the scientists who produce that published research, at least in terms of methodological rigor. We still get bad scientists; it's just that institutions won't allow them to publish their worst work. This is doubly troubling if we then expect those same corner-cutting researchers to perform exemplary peer review.

We next turned to an exploration of funding strategies. We first studied very simple strategies and found that a strategy of purely random funding allocation is little better than directly funding labs based on publication history. We did find that if funding agencies could ef-

fectively target those research groups using the most rigorous methods, the degradation of research quality can be completely mitigated. This is, however, a big "if." Rigor is notoriously difficult to assess, and it is probably quite unrealistic to assume that funders could consistently and accurately infer the quality of a lab's methods. So it appears at first glance that random allocation is unhelpful and that funding focused on rigor works but is probably a pipe dream.

These results were discouraging, to say the least. However, we then started paying more attention to the emerging literature on *modified funding lotteries*, which incorporate aspects of funding strategies focused on both randomness and rigor. Recently, a number of scholars and organizations have supported a type of lottery system for allocating research funds (Barnett 2016; Fang and Casadevall 2016; Bishop 2018; Avin 2018; Gross and Bergstrom 2019), usually proposing that a baseline threshold for quality must first be met in order to qualify projects for consideration in the lottery. Although rigor may be difficult to assess precisely, at least *some* information about the integrity of a research lab is often available. Such lotteries may confer advantages not directly related to reproducibility, including (1) promoting a more efficient allocation of researchers' time (Gross and Bergstrom 2019); (2) increasing the funding of innovative, high-risk/high-reward research (Fang and Casadevall 2016; Avin 2018); and (3) reducing gender and racial bias in funding, as well as systemic biases arising from repeat reviewers or proposers coming from elite institutions (Fang and Casadevall 2016). Such biases can lead to cascading successes that increase the funding disparity between those who, through luck, have early successes and those who don't (Bol, de Vaan, and van de Rijt 2018). However, the potential influence of modified lotteries on reproducibility had not previously been studied.

We investigated a funding strategy in which funds were awarded randomly to the pool of *qualified* applicants. Applicants were qualified if their methodological rigor (equivalent to the inverse of their characteristic false positive rate) did not fall below a threshold. We found that this strategy could be extremely effective at reducing false discoveries, even when using fairly modest thresholds (such as restricting funding to labs with false positive rates below 30%). Even better, when modified lotteries were paired with improvements to peer review and publication bias, the model produced dramatic improvements to both the scientific literature *and* the scientists producing that literature. This indicates that funders who prioritize research integrity over innovation or productivity may be able to exert a positive influence over the landscape of scientific research above and beyond the individual labs they fund.

Many of the interventions heralded by the Open Science movement—including registered reports, preprints, open data, and the like—have undeniable value. This model indicates that these interventions are likely to be insufficient to sustain the persistence of high-quality research methods as long as there are strong incentives for maximizing simple quantitative metrics like publication quantity and impact factor, which act as proxies for desirable but complex and multifaceted traits. On the other hand, the model also provides room for cautious optimism. Even in the face of strong selective pressures for publication at the key bottlenecks of hiring and promotion, science may nevertheless be improved by countervailing pressures at other bottlenecks, such as the competition for funding, if they promote rigor at the cost of productivity.

Discussion

This is a chapter about how institutional incentives shape behavior in academic science. Methods are shaped by cultural selection for practices that help researchers optimize the criteria on which they are judged, hired, and promoted. Selection can shape practices even in the absence of strategic behavior to change those practices. If methods are heritable, selection is sufficient to be damaging. The improvements promoted by the Open Science movement, as well as by well-intentioned funding agencies, are important. The models indicate that they can do some good. Beyond what is captured by the models, these practices may produce normative shifts by becoming associated with prestige and by promoting the informal punishment of transgressors. However, the models also indicate that Open Science practices are not sufficient if selection continues to favor easily measured evaluation metrics over more holistic, multidimensional assessments of quality. This conclusion forces us to consider exactly what properties we want in our academic scientists.

This is also a chapter about cultural evolution. In the last few decades, a new interdisciplinary field has emerged. It has provided formal models, increasingly backed by empirical research, of how individuals maintain cooperative participation (e.g., Boyd and Richerson 1992; Hooper, Kaplan, and Boone 2010), how they acquire and transmit cultural information (e.g., Henrich and Gil-White 2001; Kendal et al. 2018), and how the population dynamics of cultural traits unfold as a result (e.g., Boyd and Richerson 1985, 2002; Mesoudi 2011; Turchin et al. 2013; Waring, Goff, and Smaldino 2017). In October 2018, the Cultural Evolution Society held its second meeting in Tempe, Arizona, with over two hundred participants representing psychology, anthropology, archaeology, behavioral ecology, genetics, linguistics, economics, sociology, engineering, and mathematics. It behooves those who are interested

in the science and sociology of science to pay attention to this field, for its primary focus is cultural stability and the dynamics of cultural change. It also appeared to me, as a participant, that much of the science presented was of unusually high quality. It is possible that, when one has to present work to those unfamiliar with the methodological norms of a small subfield, there is a strong incentive to be extraordinarily thorough and transparent. Although field-specific expertise is invaluable in assessing research, it may also be that cross-disciplinary communication has an important role to play in maintaining methodologically rigorous research.

This is also a chapter about models. I have presented a series of five models, each of increasing complexity, to help us understand and explain the process and cultural phenomenon of scientific research. How we model science shapes our ability to identify both problems and solutions. Even at their most complex, models involve drastic oversimplification. The models I have presented focus on hypothesis testing—the fact-finding portion of science—and ignore the critical role of theory building. In these models, hypotheses are independent of one another, rather than interconnected. Hypotheses are formulated as clearly true or false, and results are formulated as unambiguously positive or negative. The later models characterize competition as being solely about publication, whereas network effects and research topics also drive success. Perhaps most importantly, the models ignore innovation and the social significance of results. Taken in isolation, these models represent a fairly crude way of thinking about science. However, the point of a model is not to capture all the nuances of a system. The point of a model is to be stupid (Smaldino 2017a). By being stupid, a model clarifies the aspects of the system we should be paying attention to and makes clear the aspects we do not include, forcing us to consider their influence on a system we now at least partially understand. Models are not the sum total of our understanding, but they can scaffold our imaginations toward a richer and deeper understanding of complex systems (Haldane 1964; Schank, May, and Joshi 2014). The models I have presented have focused on the factors that make positive results more or less likely to represent true facts. That is an important question about how science works, but it is far from the only question. A more complete understanding of the system requires many models with many perspectives and many different stupid oversimplifications. With them we can consider, for example, how false facts are canonized through publication bias (Nissen et al. 2016; Romero 2016), how funding allocation affects the efficiency of research effort (Avin 2018; Gross and Bergstrom 2019), how group loyalties and gatekeeping institutions can stifle innovative paradigms

(Akerlof and Michaillat 2018), how scientists select important research questions (Strevens 2003; Weisberg and Muldoon 2009; Thoma 2015; Alexander, Himmelreich, and Thompson 2015; Bergstrom, Foster, and Song 2016; O'Connor 2019; Zollman 2018), and how we might develop better theories (Stewart and Plotkin 2021).

To some extent, this is a chapter about how incentives for publication ruin everything and how those incentives have to change. However, it should not be taken as a story about how we academics are powerless in the face of the mighty incentives. It's true that we inherit the culture into which we are born and develop, but it's also true that we collectively create the culture in which we participate. Collectively, we have the power to change that culture.