History, Teaching, and Public Awareness

Why Is It so Hard to Do Good Science?

Ray Dingledine[1]

DOI:http://dx.doi.org/10.1523/ENEURO.0188-18.2018

[1]Department of Pharmacology, Emory University School of Medicine, Atlanta, GA 30322

Abstract

"Good science" means answering important questions convincingly, a challenging endeavor under the best of circumstances. Our inability to replicate many biomedical studies has been the subject of numerous commentaries both in the scientific and lay press. In response, statistics has re-emerged as a necessary tool to improve the objectivity of study conclusions. However, psychological aspects of decision making introduce preconceived preferences into scientific judgment that cannot be eliminated by any statistical method. The psychology of decision making, expounded by Kahneman, Tversky, and Thaler, is well known in the field of economics, but the underlying concepts of cognitive psychology are also relevant to scientific judgments. I repeated experiments carried out on undergraduates by Kahneman and colleagues four to five decades ago, but with scientists, and obtained essentially the same results. The experiments were in the form of written reactions to scenarios, and participants were scientists at all career stages. The findings reinforce the roles that two inherent intuitions play in scientific decision making: our drive to create a coherent narrative from new data regardless of its quality or relevance and our inclination to seek patterns in data whether they exist or not. Moreover, we do not always consider how likely a result is regardless of its *p* value. Low statistical power and inattention to principles underpinning Bayesian statistics reduce experimental rigor, but mitigating skills can be learned. Overcoming our natural human tendency to make quick decisions and jump to conclusions is a deeper obstacle to doing good science; this too can be learned.

Key words: Bayesian; cognitive psychology; decision making; premortem; reproducibility; rigor

Significance Statement

Societal approaches to improving the rigor and reproducibility of preclinical biomedical science have largely been technical in nature with a renewed focus on the role of statistics in good experimental designs. By contrast, the importance of preconceived notions introduced by our very human nature has been underappreciated for their influence on scientific judgments. Explicitly recognizing and addressing these cognitive biases, and including such strategies as carrying out a "premortem" before embarking on new experimental directions, should improve scientific judgments and thereby improve the quality of published findings, eventually boosting public confidence in science.

Introduction

Most failures in advanced (phase 3) clinical trials for small molecule drug candidates can be traced to insufficient efficacy (Hay et al., 2014). Although clinical trial design fettered by corporate needs and the meager predictive value of many preclinical animal models undoubt-

Received May 17, 2018; accepted July 8, 2018; First published September 4, 2018.

The author declares no competing financial interests.

Author contributions: R.J.D. designed research; R.J.D. performed research; R.J.D. analyzed data; R.J.D. wrote the paper.

This study was partially funded by the National Institute of Neurological Disorders and Stroke Grant R01NS097776.

Acknowledgements: I thank discussions with my lab group, Lance Waller and Sherry Edwards, and comments on a late draft by Linda Boland.

Correspondence should be addressed to Ray Dingledine, Ph.D., Department of Pharmacology, Emory University School of Medicine, 1510 Clifton Road, Atlanta, GA 30322, E-mail: rdingle@emory.edu.

DOI:http://dx.doi.org/10.1523/ENEURO.0188-18.2018

edly contribute to these failures, especially for some neuropsychiatric disorders, there is a growing realization that suboptimum experimental practices in preclinical research are also a prominent cause of the problem. Overconfidence in favorable *p* values, low statistical power, and the waning of traditional good experimental design practices such as blinding and inclusion of both positive and negative controls have all been highlighted as likely sources of misleading preclinical research (Ioannidis, 2005, 2018; Begley and Ellis, 2012; Button et al., 2013; Marino, 2014; Halsey et al., 2015; Steward, 2016; Colquhoun, 2017). Recent articles in the lay press make the general public aware of this issue (Leaf, 2013; The Economist, 2013, 2018; Carroll, 2017). Poor reproducibility of scientific research undermines public confidence in science and leads to waste of resources when investigators attempt to replicate and extend fallacious findings.

The scientific community is responding vigorously to this growing "crisis of confidence" in the reliability of preclinical science. Universities are reemphasizing a working knowledge of statistics that extends beyond "plug and play" software. Professional societies such as the Society for Neuroscience are providing web-based resources for enhancing the rigor and transparency of research (https://www.sfn.org/news-and-calendar/news-and-calendar/news/professional-development/virtual-conference-scientific-rigor-transparency-science). Journals are providing checklists meant to enhance transparency and, in some cases, employing statisticians on their editorial boards to check the veracity of findings. A checklist has been introduced to help recognize research that might have one or more procedural flaws (Begley, 2013). Finally, the National Institutes of Health has introduced mandatory sections in grant proposals that are meant to improve rigor (https://grants.nih.gov/reproducibility/index.htm). This is all good and helpful, but renewed attention to statistics and good experimental design practices will not, by themselves, solve this problem. Inherent in the goal of improving the reproducibility of preclinical science is the need to change behavior. Accepting the value of statistics or proper experimental design is rarely enough to bring about the desired change. In this regard, science can benefit from insights gleaned in the fields of investing, finances and marketing from considerations of how people arrive at judgments and make decisions.

In the late 1960s, two young psychologists, Daniel Kahneman and Amos Tversky, initiated a series of experiments designed to understand how people make decisions when presented with fragmentary information of uncertain relevance. Thus seeded the field that evolved into behavioral economics (Thaler, 2016). Their experiments took the form of questions that subjects would answer, and a comparison of their answers with statistically-apt outcomes. Among their major findings was that preconceived notions and unconscious emotions often dominate decision making when one is presented with new or unfamiliar data (Tversky and Kahneman, 1974; Kahneman and Tversky, 1979). In their explanation, the nature of decision making often involves over-reliance on intuition, to a degree not explicitly recognized, and underappreciation of the role of chance in events. Kahneman (2011) describes two approaches to decision making, unconscious, intuitive and reflexive, or conscious, deliberate and reflective, which he terms fast and slow thinking, respectively. Fast thinking is easily influenced by cognitive biases, whereas slow thinking is more resistant. A major outcome of their work is the realization that emotion-driven preconceptions and hopes play a large role in financial, investing, political, and economic decisions, for example, the notions that we are all good investors (Belsky, 2016), that trade wars are quickly winnable (North, 2017), or in the present context, that our scientific projects will proceed without insurmountable hitches.

Findings from cognitive psychology have been applied to fields as disparate as hostage negotiation (Voss, 2016) and healthcare system planning (Harrison et al., 2018). Science prides itself on objective thinking; I wondered whether insights gleaned from cognitive psychology were relevant to decisions made by scientists. Kahneman and Tversky's subjects were mostly undergraduate students. My notion, and the hypothesis I tested, was that a scientifically literate population would respond more objectively to the survey questions originally posed in the 1960s and 1970s. I was wrong.

A three-page questionnaire consisting of five questions was designed based on the projects originally conducted by Kahneman and Tversky as described below. The surveys were distributed to research personnel in three basic science departments, consisting of faculty, postdoctoral fellows, graduate students, and senior research technicians. Adult English-speaking scientists were the intended target population without regard for gender or ethnicity. Voluntary participants then completed the survey anonymously, estimated to require $<$15 min, and the surveys were returned to the author for analysis. There were 44 respondents out of ~70 surveys distributed. The protocol was approved by Emory's IRB.

Prologue

I began this project by seeking the opinions of colleagues about difficulties they encounter in their day-to-day pursuit of biomedical research. When asked what, aside from the scientific project itself, challenges them the most in their research, a sampling of postdoctoral fellows, students, and research technicians had several reactions:

- The PI can become distant from the project; expectations and reality diverge.
- Lab members without sufficient training can be expected to work on the project.
- Reproducing a method developed in another lab can be difficult.
- Getting all parameters of an experiment under control can be problematic.
- Perverse publishing incentives are rampant – quantity and impact factor over quality.
- Appropriate controls can sometimes be difficult to identify.
- Rigorous experiments can be labor intensive and expensive.

Regarding the last point, it is helpful to separate experiments into two categories (cf. Kimmelman et al., 2014). Exploratory experiments are quick and sometimes without all necessary controls, are often underpowered, done to explore and adjust experimental parameters to sharpen the hypothesis, and do not usually incorporate a blinding step in analysis. Low statistical power and the absence of a blinding step result in many false positives and findings that inevitably regress toward the mean during in-house replication attempts. The large majority of preclinical research falls into this category, and much of biomedical graduate training emphasizes this approach. On the positive side, experimental results from disparate approaches that test the same hypothesis, each of which is underpowered statistically, can be combined to produce a coherent conclusion. In the absence of a blinding step, however, unconscious bias still reduces confidence. Definitive experiments, by contrast, require the experimenter and/or data analyzer to be unaware of treatment groups, involve a larger sample size and carefully restrict the number of comparisons to be made. Often the results and conclusions will form the basis of a larger effort (e.g., a multi-year grant application, a Ph.D. thesis project, or a clinical trial). Definitive experiments can be expensive, both in time and money, but the results have a better chance of standing the test of time. Most of what follows is applicable to both categories of experiment.

The informal responses in the bulleted list above indicate scientists are well aware of logistic impediments to doing good science. No one, however, raised the possibility that cognitive biases might present difficulties. I repeated several experiments that Kahneman and Tversky conducted on undergraduates in the 1960s and 1970s to determine if objectivity is enhanced in a scientifically literate population. A set of five survey questions that had been originally designed by Kahneman and Tversky was distributed to colleagues in several basic science departments at Emory University School of Medicine, and the results tabulated. Responses to one of the five questions suggested that its wording was ambiguous, so what follows describes responses to four survey questions. The questions are presented verbatim below.

The Law of Small Numbers

One survey question asked respondents to consider the following:

> "A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50 percent of all babies are boys, although the exact percentage varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower. For a period of 1 year, each hospital recorded the days on which more than 60 percent of the babies born were boys. Which hospital do you think recorded more such days?"

Forty-four respondents answered as follows:

The larger hospital: 3
The smaller hospital: 18
The hospitals are within 5% of each other: 23.

Approximately 59% of scientists thought the large and small hospitals would be similar or the larger hospital would have more such days. Sampling theory, however, indicates that the smaller hospital would have about twice as many days in which >60% of infants were boys (98 ± 4 vs 48 ± 2 d with $n = 10$ repetitions of a bootstrapped model). Interestingly, 41% of today's scientists chose the correct answer compared to only 20% of the original sample of 97 undergraduates with little or no background in probability or statistics. The *p* value of 0.041 comparing the two cohorts (Fisher's exact test) is encouraging of progress until one calculates the statistical power of the comparison, which is only 52% (G∗power 3.1.9.2). Low statistical power increases the likelihood of a false positive (Steward, 2016).

In a second question, a group of faculty, postdocs, and graduate students reacted to the following scenario:

> "The distribution of newly-diagnosed kidney cancer in the 3,141 counties in the U.S. reveals that the counties with lowest incidence are mostly rural, sparsely populated, and located in traditional Republican states in the Midwest, the South and the West. What do you make of this?"

Non-political answers were (with number of respondents in parentheses if >1):

> Better diet (5; eating more greens, less processed foods, organic); poor diagnosis in rural areas (3); lower anxiety and stress (3); rare forms of cancer less likely if fewer people (2); healthier lifestyle; less medical care; early mortality due to other causes; counties are undersampled; smaller population leads to lower numbers with cancer; smaller sample size means greater variance.

A separate group was presented with the same scenario, except the word "lowest" was replaced by "highest." Their responses:

> Low access to health care or preventive care (7); pesticides or other contaminants in groundwater are carcinogenic (6); increased recent screening (3); low access to healthy foods (2); socioeconomic status leads to higher stress; education level; ethnicity; less healthy lifestyle.

It is interesting that stress, diet, and lifestyle feature prominently in both scenarios but are used to reach opposite conclusions. Both scenario statements are correct; moreover, a county that has a high incidence one year can have low incidence the next. The two words in both scenarios most relevant to understanding the nature of the problem are "sparsely populated," as shown in Figure 1; low *n* leads to high variability.

Both of these examples demonstrate an inherent belief in the Law of Small Numbers (Tversky and Kahneman, 1971), which is the tendency to accept data from a small sample as representative of the whole population. This inclination likely contributes to the willingness to do underpowered experiments and accept the outcome if one is lucky enough to have a favorable *p* value. These data are congruent with the proposal (Kahneman, 2011) that our tendency to build a narrative around new data, however irrelevant, is strong and can prevent us from getting at the underlying drivers of our observations.

Intuitive Pattern Seeking

People seek patterns, as illustrated by the following examples. A group of scientists were asked to consider three possible sequences of the gender of eight infants born in a row at the same hospital:

BBBBGGGG
GGGGGGGG
BGBBGBGB

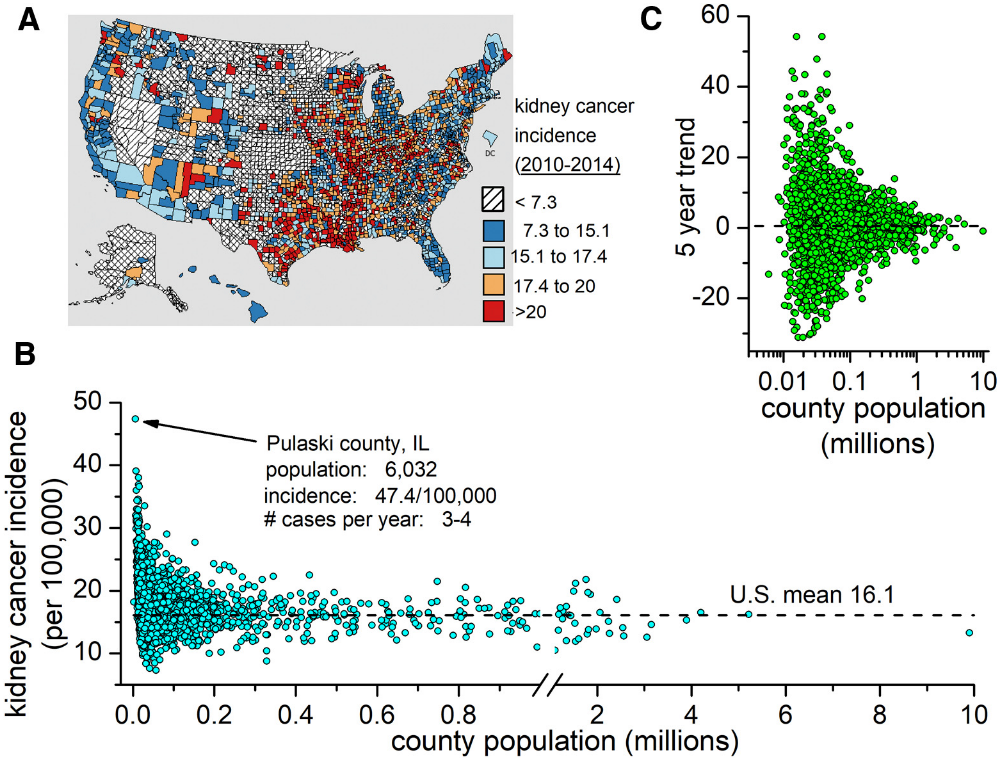

Figure 1. Low sample size produces high variance. ***A***, Map of the average kidney cancer incidence in each of the 3141 counties of the United States between 2010 and 2014. ***B***, Plot of incidence as a function of county population. ***C***, Five-year trend in incidence versus population. Data from the CDC National Program of Cancer Registries.

They were asked: "Do these sequences seem equally likely?"

Yes: 20 respondants

No: 24 respondants

Any specific sequence is as likely as any other, but 55% of respondents thought otherwise. Their written comments indicated that many thought eight girls in a row was extremely unlikely, which is true (probability = 2E-8 or 0.4%) but no less likely than the specific sequence below it. We all try to look for patterns in our data, whether they exist or not. Random events can produce runs of seeming regularity (Taleb, 2004), and our pattern-seeking human nature can easily reach a false conclusion. It can be particularly difficult to trust the eye when evaluating large datasets, because seemingly small differences in analysis procedure can result in substantial visual differences in patterns, as illustrated by different clustering algorithms of the same microarray dataset in Figure 2. In Figure 2*A*, the data appear to segregate into two main groups, whereas in Figure 2*B*, the same data collect into six groups.

A final, somewhat treacherous, example of pattern seeking is illustrated by the spatial pattern of bombs dropped on London during the eight-month-long blitz in 1940–1941. Although when viewed from a distance the bombing pattern appears very dense (Fig. 3, left), a detailed view of smaller regions reveals scattered neighborhoods that were spared throughout the eight-month period (Fig. 3, right). Reports at the time speculated that the bombing pattern was not random (Feller, 1950, p 120) and that German spies might have resided in the spared neighborhoods (Kahneman, 2011, p 116), a conclusion that could only have been socially disruptive during a trying time. However, a binomial analysis of the spatial pattern of impact sites was fully consistent with a random pattern of bomb locations ($p = 0.88$ by χ^2; Clarke, 1946). Taken together, these examples support a tendency of scientists and non-scientists alike to find patterns in data.

Ignoring the Likelihood of a New Result

The base rate is the probability that a member of a specific population will have a certain characteristic, assuming that we know nothing else about this individual. Does the base rate influence decisions when new data are introduced? A group of 43 scientists considered the following scenario:

> Chris is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to have little feel and little sympathy for other people, and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense.

Respondents were asked to rank order nine fields by the likelihood that Chris is in that field (1 = most likely, 9 = least likely). Forty-three scientists replied (Table 1).

There is no right or wrong answer to this puzzle of course, but even a cursory perusal of the results suggests that the respondents paid little attention to, or were unaware of, the base rates of employment; the four fields deemed least likely have, together, approximately six-fold

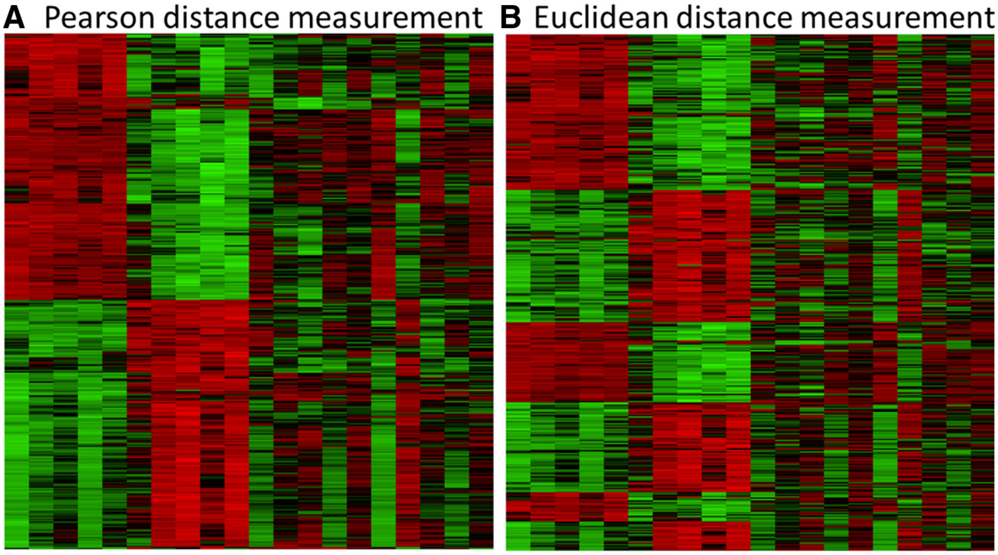

Figure 2. Microarray clustering procedure drastically alters the visual pattern. The data were from dentate granule cells provided by seven laboratories and consist of median log2 expression values of 398 genes that were differentially expressed (FDR $<$ 0.05) with $\geq$2-fold expression change between control rats and rats in three status epilepticus (SE) models at different times after SE (Dingledine et al., 2017). Complete linkage and row (gene)-clustering only; the relative positions of columns (treatment groups) in ***A***, ***B*** are unvarying. The only difference in the two procedures was the distance measurement used. The Heatmapper tool was used for clustering and visualization (http://www1.heatmapper.ca/expression/).

higher employment than the four deemed most likely. A full 67% of respondents felt that Chris was more likely to be in library science than in business administration, a field that has 64-fold higher employment. Interestingly, the rank order of fields was nearly identical to that found when the study was originally done, although the responding cohorts were as different as undergraduates and scientists, and the tests were separated by 45 years (Fig. 4; Kahneman and Tversky, 1973).

A related problem arises when one does a properly designed experiment and obtains a result with a statistically significant *p* value. The minimalist conclusion is that the null hypothesis is unlikely to be true, although what we are actually interested in is whether our hypothesis is correct. These are two very different statements. Falsifying the null hypothesis does not imply that our hypothesis is correct; other explanations might serve better. What determines whether a statistically significant result is ac-

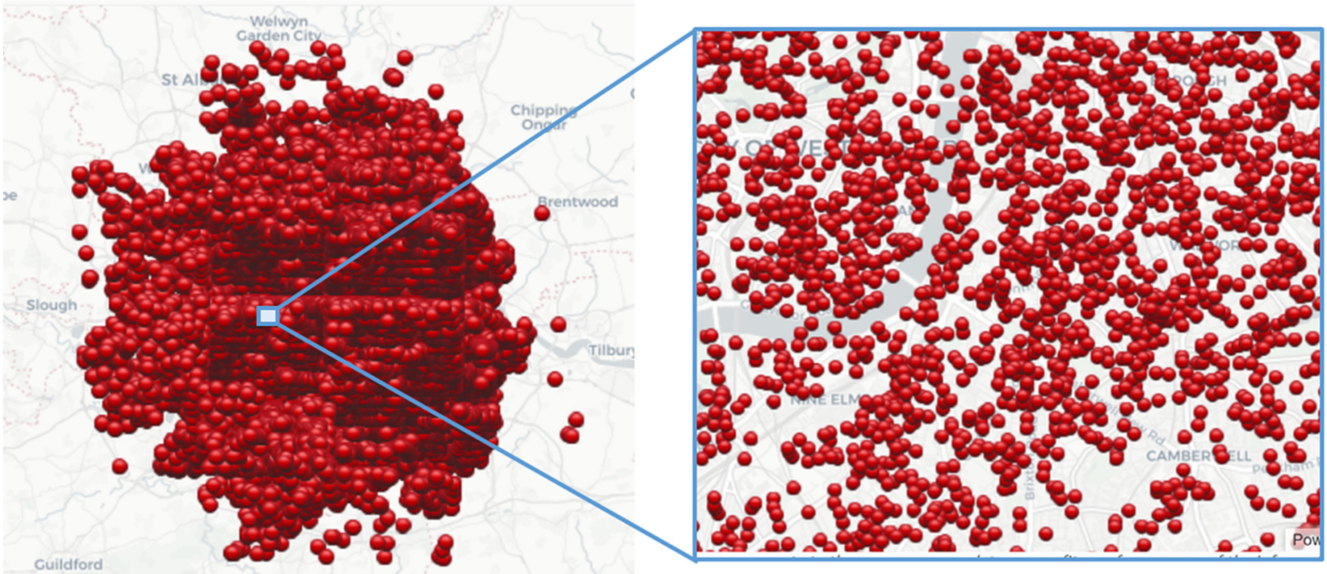

Figure 3. Pockets of seeming regularity in a random process. The left panel shows the location of each of the bombs dropped on London during the eight-month blitzkrieg of 1940–1941. Although the city was blanket-bombed, there were isolated neighborhood-size areas in which no bombs fell as shown on the right panel (from http://www.dailymail.co.uk/sciencetech/article-2243951/The-astonishing-interactive-map-EVERY-bomb-dropped-London-Blitz.html).

Table 1. Failure to take employment numbers into account when guessing professional field.

2018 Rank	Mean score	Field	2016 United States employment*
1	2.7	Computer science	4,165,140
2	3.0	Engineering	2,499,050
3	3.5	Physics or biology	150,970
4	4.4	Library science	222,760
5	4.8	Law	1,075,520
6	5.5	Business administration	14,371,980
7	6.3	Healthcare	12,361,980
8	7.0	Humanities and education	12,928,630
9	7.9	Social sciences and social work	2,264,070

*Bureau of Labor Statistics.

tually correct rather than being a false positive? Traditional Fisherian statistics highlights the role of statistical power and α (the probability of observing that result or more extreme results, assuming the null hypothesis is correct) in estimating the probability that a statistically significant finding is actually correct. However, the thought experiment presented in Figure 5 indicates that α seriously underestimates the proportion of false positive findings which, in the three examples shown, ranges from 20 to 60% as explained below.

We begin by considering an experiment that tests 1000 hypotheses, 100 of which are expected to be true based on our experience. If our experimental design yields 80% power, only 80 of these true hypotheses will be identified as such and the remaining 20 will be missed, i.e., false negatives. With $\alpha = 0.05$, the remaining 900 hypotheses are also separated into two groups, with 45 ($=0.05*900$) being false positives and the remaining 855 appropriately identified as "no effect" or true negatives. These predictions are distributed into the outcomes table in Figure 5*A*. Given this situation, the probability that any single hypothesis that tests positive is actually correct is $80/(80 + 45) = 64\%$, a far cry from the 95% that is often mistakenly assumed with $\alpha = 0.05$. Thus, the probability of a false positive is 36% in this case, not 5%. Going on, if our experimental design only allowed a statistical power of 30%, as is common in many published neuroscience papers (Button et al., 2013), the predicted outcomes table indicates that the chance of a single measured positive result actually being true is reduced to only 40% (Fig. 5*B*), i.e., a measured positive will be a false positive 60% of the time. Now the critical part. Assume that with increased knowledge of our experimental system we test hypotheses in a restricted space such that now 40% of the hypotheses are expected to be true (Fig. 5*C*). Under these conditions, our chance of correctly identifying a single hypothesis that measured true rises to 80%, i.e., the false positive rate is reduced to 20%, although α and statistical power are unchanged.

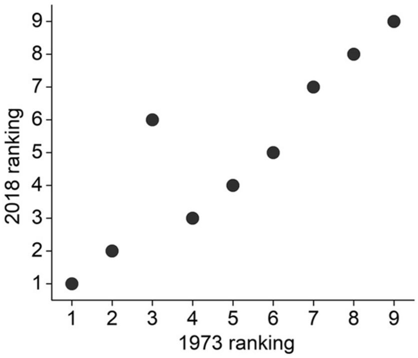

Figure 4. Similarity of judgments in populations of undergraduates and scientists separated by 45 years.

Thus, the *p* value, by itself, does not provide good evidence favoring any non-null hypothesis. Your chance of correctly interpreting a statistically significant result depends on α, statistical power, and *the probability of the result based on prior knowledge* (Fig. 5*D*). Similar arguments have been made by Colquhoun (2014) and Button (2016), among others. The prior knowledge component built into a Bayesian framework can build formal confidence in an hypothesis as results from successive experiments are added (Goodman, 1999; van de Schoot et al., 2014; Buchinsky and Chadha, 2017), always recognizing, however, that an hypothesis, no matter how many supportive repetitions exist, can be refuted by a single well-conceived negative finding (Popper, 1959). The major problem with a Bayesian approach is that it is often difficult to estimate the prior probability of a result before the experiment is conducted, although workarounds have been suggested (Matthews, 2001; Held, 2013; Colquhoun, 2017). A proposed alternative to a Bayesian approach would be to accept a result as statistically significant only if $p < 0.005$ rather than $p < 0.05$ (Benjamin et al.,2018; Ioannidis, 2018); this would reduce the false positive rates in Figure 5 to 5.3%, 13%, and 2.4% in panels *A*, *B*, and *C*, respectively, which are more palatable than the typical situations with $p < 0.05$.

Conclusions and Recommendations

Why is it so hard to do good science? The work of Kahneman, Tversky, and Thaler, supported by the recent tests described here, points out that a major part of the problem is rooted in our human nature, which makes us prone to jump to conclusions. We have tendencies to favor the law of small numbers, to spin narratives out of datasets of questionable relevance, and to seek patterns in noisy data. Kahneman and Tversky's biases are prominent in the best of us: it is easier, more social, and more comfortable to think intuitively and quickly, whereas rational, contemplative thinking is difficult, can be tiring, and is frequently a solitary exercise. Second, we often have too much confidence in a favorable *p* value, and sometimes do not pay enough attention to whether a new result is likely to be true regardless of statistical significance. One pernicious outcome of this situation is the proclivity for some journals to publish "very unexpected" findings that nonetheless have statistical significance.

The experiments and scenarios described here are relevant to how people evaluate tests of hypotheses, not how we form hypotheses. The critical thinking skills required to rigorously test hypotheses are quite distinct from the cognitive processes involved in the initial generation of hypothe-

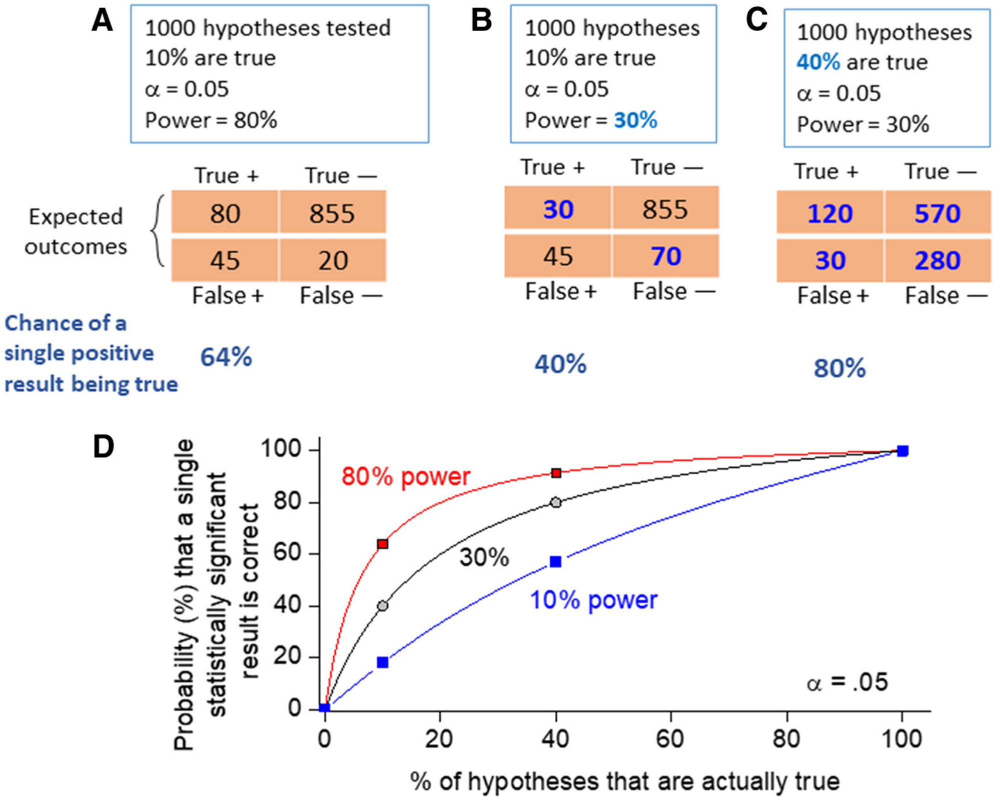

Figure 5. Demonstration of the value of a Bayesian approach when interpreting a single experiment. ***A***, Consider the scientist facing the challenge of testing 1000 hypotheses, for example, in a proteomics or drug screening experiment, in which each protein measured or compound assayed tests the hypothesis that the protein expression is changed or the compound is active. Without knowledge of the target class only a small fraction of these hypotheses (say 10%) are true. If the power of the experiment is 80%, 80 hypotheses will be correctly identified as true (marked "true +") and 20 will be false negatives ("false –") as shown in the expected outcomes table. Of the 900 remaining false hypotheses, 45 will be incorrectly ascribed as being true ("false +" in the table) with $\alpha = 0.05$, and 855 will be correctly ascertained as negatives ("true –"). Thus, the chance of correctly interpreting a positive outcome (equivalent to precision in a ROC analysis) is only 64% (=80/125). ***B***, If the statistical power of the experiment is 30%, the chance of correctly interpreting a positive result drops to only 40% (=45/75). Blue in the outcomes table represent numbers that have changed from the previous panel. ***C***, Now imagine that, with experience, one is working in a more restricted target space in which 40% of hypotheses are expected to be true: the chance of correctly interpreting a positive result jumps to 80% even when power = 30% (=120/150). ***D***, A plot of these variables (plus the 10% power case) shows that the chance of correctly interpreting a single outcome depends on α, power, and the a priori probability that the hypothesis is true.

ses. It is striking that answers to the survey questions were so similar regardless of whether respondents were undergraduates in the 1960s and 1970s, or scientific investigators engaged this year. Admittedly, the outcome of the experiments reported here only show that scientists exhibit human features and do not directly demonstrate that we are subject to cognitive biases when planning and evaluating scientific projects. This could be the subject of future work that addresses the circumstances under which scientific decisions are influenced, although senior scientists I have spoken with recognize the relevance to their own research.

Deciding when it is allowable to exclude an outlier, how to incorporate both positive and negative controls, how to determine the number of subjects that will provide an appropriate level of confidence in one's findings, and selection of the proper statistical test, are all important technical skills that can be learned. Circumvention of cognitive biases, however, is also important in most experiments but particularly preclinical animal trials, cell culture experiments, any experiment in which outcomes are compared among two or more experimental groups.

Notwithstanding the above, plenty of excellent science that stands the test of time is reported by reputable journals every month. How can the substantive investigator improve the quality of scientific endeavors, develop more confidence in our findings, and guard against a premature exit from a research career? First and probably foremost, it is helpful to recognize explicitly the human nature features that can lead us astray; we can then attempt to use Kahneman's "slow thinking" strategy, recognizing and trying to avoid common biases that unhelpfully support our preconceived notions. Second, we could more often employ good experimental practice that avoids the red flags identified by Begley (2013). It can be helpful in this regard to openly separate experiments into exploratory and definitive categories, and to commit to performing a definitive experiment once the parameters are sufficiently well understood. Of course, there is a need to balance this suggestion against resource availability and time deadlines. Incorporating a Bayesian framework into experimental design could also allow quantification of confidence in the outcome of several small experiments that are considered together. With a Fisherian outlook, accepting a result as "statistically significant" only if $p < 0.005$ will also reduce false positives.

Finally, before embarking on a new line of experiments it can be helpful to perform a "premortem," i.e., assume the experiment failed and list all potential reasons (cf. Klein, 2007). When setting out in an unfamiliar direction, it is too easy to overestimate the benefits and underestimate the chance of failure with its associated costs. A premortem can present the opportunity to improve rather than autopsy a project. During a premortem, a group of participating scientists would likely identify a wider range of potential problems than just one or two scientists. Framing the discussion by assuming the experiment failed can provide a means for the thoughtful yet somewhat timid members of a group to voice their concerns. I have found that this exercise often encourages the addition of new controls that help us understand why the initial experimental outcome was not as hoped, which in turn accelerates progress toward eventual success.

All of these suggestions would have us acting more like the ancient Greek Titan, Prometheus, and less like his brother, Epimetheus. Prometheus has come to symbolize "forethought" and effective planning, with Epimetheus representing "afterthought." Recall that in the Greek legends, Prometheus helped advance mankind in many ways, whereas Epimetheus, in a rush, married Pandora and helped open her box.

References

Begley CG, Ellis LM (2012) Drug development: raise standards for preclinical cancer research. Nature 483:531–533. CrossRef Medline

Begley CG (2013) Six red flags for suspect work. Nature 497:433–434. CrossRef Medline

Belsky G (2016) Why we think we're better investors than we are. The New York Times, March 25.

Benjamin DJ, Berger JO, Johnson VE, Nosek BA, Wagenmakers EJ, Berk R, Bollen KA, Brembs B, Brown L, Camerer C, Cesarini D, Chambers CD, Clyde M, Cook TD, De Boeck P, Dienes Z, Dreber A, Easwaran K, Efferson C, Fehr E, et al. (2018) Redefine statistical significance. Nat Hum Behav 2:6–10. CrossRef

Buchinsky FJ, Chadha NK (2017) To P or not to P: backing Bayesian statistics. Otolaryngol Head Neck Surg 157:915–918. CrossRef Medline

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci [Erratum (2013) 14:451] 14:365–376. CrossRef Medline

Button KS (2016) Statistical rigor and the perils of chance. eNeuro 3.

Carroll AE (2017) Science needs a solution for the temptation of positive results. The New York Times, May 29.

Clarke RD (1946) An application of the Poisson distribution. J Ins Act 72:481–482.

Colquhoun D (2014) An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci 1:140216. CrossRef

Colquhoun D (2017) The reproducibility of research and the misinterpretation of p-values. R Soc Open Sci 4:171085. CrossRef

Dingledine R, Coulter DA, Fritsch B, Gorter JA, Lelutiu N, McNamara J, Nadler JV, Pitkänen A, Rogawski MA, Skene P, Sloviter RS, Wang Y, Wadman WJ, Wasterlain C, Roopra A (2017) Transcriptional profile of hippocampal dentate granule cells in four rat epilepsy models. Sci Data 4:170061. CrossRef

Feller W (1950) An introduction to probability theory and its applications, Vol1, p 120. London: Wiley Inc, Chapman and Hall Ltd.

Goodman SN (1999) Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med 130:1005–1013. Medline

Halsey LG, Curran-Everett D, Vowler SL, Drummond GB (2015) The fickle P value generates irreproducible results. Nat Methods 12: 179–185. CrossRef Medline

Harrison O, Kelleher T, Lopez D, Volpp K, Wright J (2018) Behavioral strategies to strengthen health programmes and policies. Geneva: World Economic Forum.

Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J (2014) Clinical development success rates for investigational drugs. Nat Biotechnol 32:40–51. CrossRef Medline

Held L (2013) Reverse-Bayes analysis of two common misinterpretations of significance tests. Clin Trials 10:236–242. CrossRef Medline

Ioannidis JP (2005) Why most published research findings are false. PLoS Med 2:e124. CrossRef Medline

Ioannidis JPA (2018) The proposal to lower p value thresholds to .005. JAMA 319:1429–1430. CrossRef Medline

Kahneman D (2011) Thinking, fast and slow. New York: Farrar, Straus and Giroux.

Kahneman D, Tversky A (1973) On the psychology of prediction. Psychol Rev 80:237–251. CrossRef

Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47:263–291. CrossRef

Kimmelman J, Mogil JS, Dirnagl U (2014) Distinguishing between exploratory and confirmatory preclinical research will improve translation. PLoS Biol 12:e1001863. CrossRef Medline

Klein G (2007) Performing a project premortem. Harvard Busn Rev 85:18–19.

Leaf C (2013) Do clinical trials work? The New York Times, July 13.

Marino MJ (2014) The use and misuse of statistical methodologies in pharmacology research. Biochem Pharmacol 87:78–92. CrossRef Medline

Matthews RAJ (2001) Why should clinicians care about Bayesian methods? J Stat Plan Inference 94:43–58. CrossRef

North P (2017) The New Yorker Collection. The Cartoon Bank Publication, 2017, January 23, p 38. Image ID: TCB-143879.

Popper K (1959) The logic of scientific discovery. London: Hutchinson & Co.

Steward O (2016) A rhumba of "R's": replication, reproducibility, rigor, robustness: what does a failure to replicate mean? eNeuro 3. CrossRef

Taleb NN (2004) Fooled by randomness: the hidden role of chance in life and in the markets. New York: Random House.

Thaler R (2016) Misbehaving: the making of behavioral economics. New York: W. W. Norton & Company.

The Economist (2013) Unreliable research. Trouble in the Lab, October 18.

The Economist (2018) Are research papers less accurate and truthful than in the past? March 17.

Tversky A, Kahneman D (1971) Belief in the law of small numbers. Psycholog. Bull 76:105–110. CrossRef

Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. Science 185:1124–1131. CrossRef Medline

van de Schoot R, Kaplan D, Denissen J, Asendorpf JB, Neyer FJ, van Aken MAG (2014) A gentle introduction to bayesian analysis: applications to developmental research. Child Dev 85:842–860. CrossRef Medline

Voss C (2016) Never split the difference. Negotiating as if your life depended on it. London: Random House Business Books.