Special Issue:
Bayesian Probability and Statistics in Management Research

Journal of Management
Vol. 41 No. 2, February 2015 421–440
DOI: 10.1177/0149206314547522

Editorial Commentary

Surrogate Science: The Idol of a Universal Method for Scientific Inference

Gerd Gigerenzer
Max Planck Institute for Human Development
Julian N. Marewski
University of Lausanne

The application of statistics to science is not a neutral act. Statistical tools have shaped and were also shaped by its objects. In the social sciences, statistical methods fundamentally changed research practice, making statistical inference its centerpiece. At the same time, textbook writers in the social sciences have transformed rivaling statistical systems into an apparently monolithic method that could be used mechanically. The idol of a universal method for scientific inference has been worshipped since the "inference revolution" of the 1950s. Because no such method has ever been found, surrogates have been created, most notably the quest for significant p *values. This form of surrogate science fosters delusions and borderline cheating and has done much harm, creating, for one, a flood of irreproducible results. Proponents of the "Bayesian revolution" should be wary of chasing yet another chimera: an apparently universal inference procedure. A better path would be to promote both an understanding of the various devices in the "statistical toolbox" and informed judgment to select among these.*

Keywords: *research methods; regression analysis; psychometrics; Bayesian methods*

> No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.
>
> —Sir Ronald A. Fisher (1956: 42)

Acknowledgments: We thank Joachim Krueger for drawing our attention to the Franco, Blau, and Zimbardo (2011) case of mindless statistics. We thank Daniela Link for comments and for the analysis of the number of p *values in the* Academy of Management Journal*. We thank François Messer, Claire Lauren Tepper, and Cvetomir Dimov for helping with this analysis.*

Corresponding author: Gerd Gigerenzer, Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition, Lentzeallee 94, 14195 Berlin, Germany.

E-mail: gigerenzer@mpib-berlin.mpg.de

> *Cognition* will require its authors to adhere to the convention that an effect *described* as "statistically significant" must have a *p*-value below .05 (for better or for worse, this is the current convention).
>
> —Gerry T. M. Altmann, editor of *Cognition* (2007: 6)

If statisticians agree on one thing, it is that scientific inference should not be made mechanically. Despite virulent disagreements on other issues, Ronald Fisher and Jerzy Neyman, two of the most influential statisticians of the 20th century, were of one voice on this matter. Good science requires both statistical tools *and* informed judgment about what model to construct, what hypotheses to test, and what tools to use. Practicing statisticians rely on a "statistical toolbox" and on their expertise to select a proper tool; social scientists, in contrast, tend to rely on a single tool. In the words of psychologist Abraham Maslow (1966: 15), "if all you have is a hammer, everything looks like a nail."

Judging by their behavior, many social scientists vote with their feet against an informed use of inferential statistics. A majority compute *p* values all the time, a minority always compute confidence intervals, and a few calculate Bayes factors, irrespective of the problem at hand. These routines are surprising, given that in most psychological experiments—unlike in election polls and quality control—researchers do not draw random samples from a population or define a population in the first place. Thus, no one knows to what population an inference actually refers.

Determining significance has become a surrogate for good research. Francis Bacon (1620/1902) used the term *worshipping idols* for a specific and pernicious sort of error. In Bacon's view, it is better to have no beliefs than to embrace falsehoods, because false idols block the way toward enlightenment. In the sciences, an idol provides a surrogate for the real object and requires from its worshippers delusions about its power (Daston, 2005).

This article is about the idol of a universal method of statistical inference. Advocated as the only game in town, it is practiced in a compulsive, mechanical way—without judging whether it makes sense or not. Here are a few illustrations.

Mindless Statistical Inference

In an Internet study on implicit theories of moral courage, participants were asked, "Do you feel there is a difference between altruism and heroism?" (Franco, Blau, & Zimbardo, 2011: 107). The far majority felt so: 2,347 respondents said yes, while 58 said no. The authors computed a chi-square test to find out whether the two numbers differed significantly, which indeed they did: "The responses indicate that there is a significant perceived difference between the ideas of heroism and altruism, $\chi^2(1) = 2178.60$, $p < .0001$" (Franco et al., 2011: 108).

In the same spirit, one of our students found that the means in his experimental and control group were exactly the same. Believing that it would be unscientific to simply report this, he was anxious to do a significance test. The *t* test revealed that the means did not differ significantly, which he dutifully reported in his thesis.

One of us reviewed an article in which the number of subjects was reported as 57. The authors calculated that the 95% confidence interval was between 47.3 and 66.7 subjects. Every figure was scrutinized in the same way, resulting in three dozen statistical tests. The only numbers with no confidence intervals or *p* values attached were the page numbers.

These cases illustrate the automatic use of a statistical procedure, reminiscent of an obsessive-compulsive disorder. Unfortunately, these are not the exception. Consider all behavioral, neuropsychological, and medical studies with humans published in 2011 in *Nature*. Eighty-nine percent of these articles reported *p* values only—without providing information about effect size, power, or model estimation. In *Science* and *Neuropsychology*, 42% and 32% did the same, respectively (Tressoldi, Giofré, Sella, & Cumming, 2013). One of our graduate students, Daniela Link, went through a top outlet, the *Academy of Management Journal*; in 2012, the estimated number of *p* values computed in an article was on average 99 (median = 89), ranging from 0 to 578.[1] Among the articles published during the 1980s in the *American Economic Review*, another flagship journal, 70% did not distinguish statistical significance from effect size, that is, economical significance. In the following decade, this percentage increased to 82% (Ziliak & McCloskey, 2004). Mindless mechanical statistics reigns.[2]

The idol of an automatic, universal method of inference, however, is not unique to *p* values or confidence intervals. As we will argue, it can invade Bayesian statistics as well. In fact, it appears to have done so already. For instance, Dennis Lindley, a leading advocate of Bayesian statistics, declared that "the only good statistics is Bayesian statistics" (1975: 106) and that "Bayesian methods are even more automatic" than Fisherian ones (1986: 6).

The Idol of a Universal Method of Inference

In this article, we make three points.

1. *There is no universal method of scientific inference but, rather, a toolbox of useful statistical methods. In the absence of a universal method, its followers worship surrogate idols, such as significant* p *values.* The inevitable gap between the ideal and its surrogate is bridged with delusions—for instance, that a *p* value of 1% indicates a 99% chance of replication. These mistaken beliefs do much harm: among others, by promoting irreproducible results.
2. *If the proclaimed "Bayesian revolution" were to take place, the danger is that the idol of a universal method might survive in a new guise, proclaiming that all uncertainty can be reduced to subjective probabilities.* And the automatic calculation of significance levels could be revived by similar routines for Bayes factors. That would turn the revolution into a re-volution—back to square one.

These first two points are not "philosophical" but have very practical consequences, because

3. *Statistical methods are not simply applied to a discipline; they change the discipline itself, and vice versa*. In the social sciences, statistical tools have changed the nature of research, making inference its major concern and degrading replication, the minimization of measurement error, and other core values to secondary importance.

Dreaming Up a Universal Method of Inference

In science and everyday life, statistical methods have changed whatever they touched (Gigerenzer, Swijtink, Porter, Daston, Beatty, & Krüger, 1989). In the 20th century, parapsychology was transformed from what was originally the study of unique messages from

the dear departed into that of repetitive card guessing. The game of baseball turned into baseball statistics: Batting averages and runs scored became its new lifeblood, with journals such as *Operations Research* publishing articles on baseball science. For centuries, medicine was based on "medical tact" in relation to the individual patient, and doctors held a long-standing antagonism toward averages. By the late 20th century, the probabilities from randomized trials began to replace doctors' intuitions, and causes were replaced by chances. Yet perhaps the most dramatic change brought about by statistics was the "probabilistic revolution": In the natural sciences, the term *statistical* began to refer to the nature of theories, not to the evaluation of data. In the social sciences, the reverse happened, with researchers dreaming up a universal method of inference—first in the form of *p* values during the "inference revolution" and nowadays in the form of Bayesian statistics as part of the proclaimed "Bayesian revolution." To understand how the latest (Bayesian) version of the old dream of a universal method of inference emerged, it is helpful to go back to these two earlier revolutions.

How Statistics Changed Theories: The Probabilistic Revolution

The probabilistic revolution upset the ideal of determinism shared by most European thinkers up to 1800 (Krüger, Daston, & Heidelberger, 1987). During the 19th and 20th centuries, it slowly made chance an indispensable part of our understanding of nature, from Darwinian variability and genetic recombination in biology to Brownian motion and radioactive decay in physics. Together with the Copernican and Einsteinian revolutions, it is one of the great scientific revolutions. Yet it differs from the others in that it did not replace any system in its own field. What it did is upset theories in other fields outside of mathematics. For instance, in physics, Newton's theory of simple cause and effect was replaced by the probabilistic causes in statistical mechanics and, eventually, by quantum theory.

Surprisingly, the social sciences inspired the probabilistic revolution in physics. When the founders of statistical mechanics, Ludwig Boltzmann and Clark Maxwell, tried to understand the erratic behavior of gas molecules, they used as their model laws guiding people, as formulated by the Belgian social statistician Adolphe Quetelet. Quetelet had discovered that human behavior, such as murder und suicide, albeit unpredictable at the individual level, could be predicted at the collective level. The collective follows statistical laws, such as normal distributions. Boltzman and Maxwell applied the same laws to the behavior of gas molecules (Gigerenzer et al., 1989).

However, although the probabilistic revolution fundamentally changed theorizing in the natural sciences, the social and medical sciences were reluctant to abandon the ideal of simple, deterministic causes (Krüger, Gigerenzer, & Morgan, 1987). With few exceptions, social theorists hesitated to think of probability as more than an error term in the equation *observation = true value + error*. Whereas Darwinian variability is necessary for selection and adaptation, measurement errors are merely a nuisance. The Harvard psychologist Edwin Boring summarized the situation in psychology as late as in 1963 in two words: "Determinism reigns" (p. 14).

Notably, probability theory shaped the social sciences differently than it did the natural sciences. It became used for *mechanizing scientists' inferences* rather than for modeling how nature works. This move promised objectivity because it replaced the subjectivity of experimenters' judgments with an automatic method. Thereby, social scientists aimed to become as

objective as natural scientists, without threatening the prevailing determinism in theories (Gigerenzer, 1987).

How Statistics Changed Methods: The Inference Revolution

The term *inference revolution* (Gigerenzer & Murray, 1987) refers to a change in scientific method that was institutionalized in psychology and in other social sciences between 1940 and 1955, first in the United States and subsequently in Europe and the rest of the world. The qualifier *inference* indicates that among all scientific tools—such as hypothesis formulation, systematic observation, descriptive statistics, minimizing measurement error, and independent replication—the inference from a sample to population grew to be considered the most crucial part of research. As stressed above, this was a stunning new emphasis, given that in most experiments, psychologists virtually never drew a random sample from a population or defined a population in the first place.

This accent on significance also pushed key scientific tools into the background. In the natural sciences, replication is a necessary part of the research process, whereas in marketing, advertising, and related areas, journal editors deemed significance to be sufficient, arguing that replication "isn't considered 'creative'" and "doesn't help get tenure or fame" (Madden, Easley, & Dunn, 1995: 79). In contrast to this dismissive attitude towards replication, W. S. Gosset of the Guinness brewery in Dublin, the father of the *t* test, considered small measurement errors to be more important than small *p* values (Gigerenzer et al., 1989). After the inference revolution, however, few researchers were concerned with true measurement errors. Simply increasing the number *N* of subjects became a surrogate for minimizing errors: The square root of *N* in the *t*-test formula appeared to do the same job. In sum, the inference revolution made inference the dominant aspect of evaluating data. Although Bayesians correctly criticize *p*-value statistics, their focus is, once again, on inference.

Science Without Statistical Inference

To understand how deeply the inference revolution changed the social sciences, it is helpful to realize that routine statistical tests, such as calculations of *p* values or other inferential statistics, are not common in the natural sciences. Moreover, they have played no role in any major discoveries in the social sciences. Consider how Isaac Newton, the father of the first unified system of modern physics, practiced research. In his *Opticks*, Newton (1704/1952) advanced propositions about the nature of light, such as that white light consists of spectral colors. Then he reported in detail a series of experiments to test these. In other words, for Newton, an experiment meant putting in place a set of conditions, derived from a theory, and then demonstrating the predicted effect. Nowhere in his writings is sampling or statistical inference mentioned. However, Newton was not hostile to statistical inference. In his role as the master of the London Royal Mint, he had to make sure that the amount of gold in the coins was neither too little nor too large. For that problem, he relied on a sampling scheme named the *Trial of the Pyx*. The trial consisted of a sample of coins drawn and placed in a box known as the Pyx (derived from the Greek term for *box*), a null hypothesis to be tested (the standard coin), a two-sided alternative, a test statistic, and a critical region (Stigler, 1999). In Newton's view, sampling and statistical inference were needed for quality control in a mint but not in science.

Similarly, inferential statistics did not contribute to Watson and Crick's discovery of the double helix or to Einstein's special relativity theory. The same can be said for virtually all of the classical discoveries in psychology. You will not find Jean Piaget calculating a *t* test. Wolfgang Köhler developed the Gestalt laws of perception, Ivan P. Pavlov the principles of classical conditioning, B. F. Skinner those of operant conditioning, George Miller his magical number seven plus minus two, and Herbert A. Simon his Nobel Prize–winning work in economics—all without calculating *p* values, confidence intervals, or Bayes factors. This was not because they did not know statistical inference. It was known long ago.

First Test of a Null Hypothesis: Infer Divine Purpose

The first known test of a null hypothesis was by John Arbuthnott in 1710. It is strikingly similar to the "null ritual" (see below) that was institutionalized in the social sciences 250 years later. Arbuthnott (1710) asked whether chance or divine planning was responsible for the slight excess of male versus female births. He observed that "the external Accidents to which Males are subject (who must seek their Food with danger) do make a great havock of them, and that this loss exceeds far that of the other Sex" (Arbuthnott, 1710: 188). To repair this loss, he argued, God brings forth more males than females, year after year. He tested this hypothesis of divine purpose against the null hypothesis of mere chance, using 82 years of birth records in London. In every year, the number of male births was larger than that of female births. Arbuthnott calculated the "expectation" of these data (D) if the hypothesis of blind chance—in modern terms, H_0— were true as $p(D|H_0) = (1/2)^{82}$. Because this probability was so small—$p < .000001$—he concluded that divine providence, not chance, rules. In his view, that empirically proved that God favors monogamy and "that Polygamy is contrary to the Law of Nature and Justice" (Arbuthnott, 1710: 189).

The first null hypothesis test impressed no one. That is not to say that statistical methods have played no role in the social sciences. Descriptive statistics have been essential for all empirical sciences, from astronomy to psychophysics, including methods of visualization and exploratory data analysis (Tukey, 1977). Descriptive statistics also provided the materials for Karl Marx and other political reformers and set the grounds for modern bureaucracy. Numbers were essential for calculating employer contributions to sickness benefits, pension funds, and other elements of the first modern welfare state in Germany (Hacking, 1987). The term *statistics* itself probably stems from the needs of states to collect demographic and economic data. Modern organizations, from Google to the National Security Agency, continue this long-standing business.

To summarize, statistical inference played little role and Bayesian inference virtually none in research before roughly 1940.[3] Among the first to promote significance testing to make their claims look more objective were parapsychologists and educational researchers (Danziger, 1990). Automatic inference was unknown before the inferential revolution, with the exception of the use of the *critical ratio* (the ratio of the obtained difference to its standard deviation). The maxim back then was "A critical ratio of three, or no PhD."

The Null Ritual

The most prominent creation of a seemingly universal inference method is the null ritual:

1. Set up a null hypothesis of "no mean difference" or "zero correlation." Do not specify the predictions of your own research hypothesis.
2. Use 5% as a convention for rejecting the null. If significant, accept your research hypothesis. Report the result as $p < .05$, $p < .01$, or $p < .001$, whichever comes next to the obtained p value.
3. Always perform this procedure.

In psychology, this ritual became institutionalized in curricula, editorials, and professional associations in the mid-1950s (Gigerenzer, 1987, 2004). In textbooks, it became dubbed "the backbone of psychological research" (Gerrig & Zimbardo, 2002: 46). Researchers are encouraged to apply this procedure automatically—by editors (see the epigram), textbook writers, and publication guidelines. For instance, the *Publication Manual of the American Psychological Association* (American Psychological Association, 1974: 19) warned its readers, "Caution: Do not infer trends from data that fail by a small margin to meet the usual levels of significance. . . . Treat the result section like an income tax return. Take what's coming to you, but no more."

Yet the null ritual does not exist in statistics proper. What does exist are conflicting theories of inference, most relevantly those of Fisher and Neyman-Pearson. Fisher, for instance, regarded Neyman's position as "childish" and "horrifying [for] the intellectual freedom of the west" (Gigerenzer et al., 1989: 105). He branded the mechanical nature of Neyman-Pearson's tests as having grown from "the phantasy of circles rather remote from scientific research" (Fisher, 1956: 100). Neyman, in turn, replied that some of Fisher's tests were "worse than useless" because their power is smaller than their alpha level (Stegmüller, 1973: 2). One rarely finds a hint at this controversy in statistics textbooks written by social scientists. As a result, the null ritual is confused with Fisher's theory of null hypothesis testing. For example, it has become common to use the term *NHST* (null hypothesis significance testing) without distinguishing between the two. But contrary to what is suggested by that misleading term, *level of significance* actually has three different meanings: (a) a mere convention, (b) the alpha level, or (c) the exact level of significance.

Three Meanings of Significance

In his early work, Fisher (1935/1971) thought of the level of significance as a mere *convention*: "It is usual and convenient for experimenters to take 5 per cent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard" (p. 13). (The reason for this convention appears to be that Fisher had tables for 5% and 1% only because his archenemy, Karl Pearson, refused to give him any others.) The convention became part of the null ritual.

But Neyman and Pearson rejected a mere convention in favor of an alpha level that required a rational scheme (Neyman, 1957). Here is a summary of their scheme:

1. Set up two statistical hypotheses, H_1 and H_2, and decide on alpha, beta, and the sample size before the experiment, based on subjective cost-benefit considerations.
2. If the data fall into the rejection region of H_1, accept H_2; otherwise accept H_1.
3. The usefulness of this procedure is limited among others to situations where there is a disjunction of hypotheses (e.g., either μ_1 or μ_2 is true), where there is repeated sampling, and where you can make meaningful cost-benefit trade-offs for choosing alpha and beta.

The alpha level is the long-term relative frequency of mistakenly rejecting hypothesis H_1 if it is true, also known as Type 1 error rate. The beta level is the long-term relative frequency of mistakenly rejecting hypothesis H_2 if it is true (also known as Type 2 error rate or 1 – power). A typical application of Neyman-Pearson testing is industrial quality control. Here, a manufacturer calculates the required sample size of products that must be examined daily to detect eventual problems in the quality of products by balancing the costs of false positives (e.g., halting the production when there is, in fact, no quality issue) and false negatives (e.g., letting low-quality products slip through).

During his long controversy with Neyman and Pearson, Fisher eventually refined his earlier position. The result was a third definition of level of significance, alongside convention and alpha level, the exact level of significance in Fisher's (1955, 1956) null hypothesis testing:

1. Set up a statistical null hypothesis. The null need *not* be a nil hypothesis (e.g., zero difference).
2. Report the exact level of significance (e.g., $p = .055$ or .045). Do not use a conventional 5% level all the time.
3. Use this procedure only if you know little about the problem at hand.

As can be seen, Fisher's mature framework differs fundamentally from the null ritual: First, one should not automatically use the same level of significance (see the epigram), and second, one should not use this procedure for all problems. Now we can also understand the muddled anatomy of the null ritual. Step 1 of the ritual—setting up only a null—stems from Fisher. But it carries the misinterpretation that null means "nil," such as a zero difference. This step contradicts Neyman-Pearson theory, where two statistical hypotheses need to be specified in order to be able to determine both alpha and beta. Power (1 – beta) is nevertheless mentioned in typical textbooks but cannot be determined in the context of the null ritual. Step 2—calculating the p value after the fact and rounding it up to $p < .05$, $p < .01$, or $p < .001$—is an odd compromise between Fisher's (1955, 1956) and Neyman-Pearson's logic that violates both. According to Neyman-Pearson, alpha needs to be determined *before* the data are obtained, and there can be only one alpha level, not three. According to Fisher, one should report the *exact* level calculated from the data but not round it up and report the rounded value as if it were a convention or an alpha level. In spite of that, scores of editors recommend reporting multiple levels of significance for one and the same analysis. For instance, the style guide for authors of the *Academy of Management Journal* ("Style Guide," 2011: 1083) explains how to report various "significance levels" (note the plural) and to award them one or more stars. Step 3 of the null ritual is the sole and unique creation of social scientists.

Bestselling Textbooks Sell a Single Method of Inference

The null ritual is an invention of statistical textbook writers in the social sciences. They became familiar with Fisher's work first, mainly through his 1935 book, and only later with Neyman-Pearson theory. After learning about Neyman-Pearson, these writers (who were mostly nonstatisticians) had a problem: How should they deal with conflicting

methods? The solution would have been to present a toolbox of different approaches, but Guilford (1942), Nunnally (1975), and many others mixed the concepts and presented the muddle as a single, universal method. Indeed, the inference revolution was not led by the leading scientists. It was spearheaded by humble nonstatisticians who composed statistical textbooks for education, psychology, and other fields and by the editors of journals who found in "significance" a simple, "objective" criterion for deciding whether or not to accept a manuscript. For instance, in 1962, the editor of the *Journal of Experimental Psychology* stated that he had "a strong reluctance to accept and publish results" not significant at the .01 level, while the significant ones were worthy of being placed in the "archives" (Melton, 1962: 553-554).

Some of the most prominent psychologists of their time vehemently objected. Stanley S. Stevens (1960: 276), the founder of modern psychophysics, complained about a "meaningless ordeal of pedantic computations." R. Duncan Luce (1988: 582), one of the architects of mathematical psychology, spoke of a "wrongheaded view about what constituted scientific progress," and Herbert A. Simon (1992: 159) made it clear that for his computer models, the "familiar tests of statistical significance are inappropriate." You will not find such statements cited in the bestselling statistical textbooks in psychology.

The idol of a universal method (in the form of the null ritual) also left no place for Bayesian statistics. Nor did some publishers. A well-known author, whose name does not matter, had a chapter on Bayesian methods in the second edition of his bestselling textbook. When asked why he deleted this chapter in all further editions, he responded that it had been done at the request of his publisher. The publisher wanted a single method that can be applied by students automatically rather than several methods that would demand discernment, because that would decrease sales. So the author acquiesced, deleting the chapter on Bayes and, along with it, the only sentence that hinted at a difference between Fisher and Neyman and Pearson. Asked about the statistical theory in which he himself believed, the author confessed that he was actually a Bayesian at heart. This case illustrates how sales, together with sacrifice of intellectual integrity, fueled the creation of the universal method of inference. In this way, textbook writers changed statistics.

Bayesianism and the New Quest for a Universal Method

> The only good statistics is Bayesian statistics. Bayesian statistics is not just another technique to be added to our repertoire . . . it is the only method that can produce sound inferences and decisions in multivariate, or any other branch of, statistics. (Lindley, 1975: 106)

> It is usual to consider the logarithm of the Bayes factor, for which the so-called "Jeffreys' scale" gives empirically calibrated levels of significance for the strength of evidence... (Trotta, 2007: 73)

Fisher and Neyman and Pearson have been victims of social scientists' desire for a single tool, a desire that produced a surrogate number for inferring what is good research. Bayes could well be the next victim. The same idol that caused the neglect of Bayesian statistics might now promote it as the only game in town. The potential danger lies in the subjective interpretation of probability, which sanctions its universal application to all situations of uncertainty.

A Short History of Bayesian Inference

The "Bayesian revolution" had a slow start. To begin with, the Reverend Thomas Bayes did not publish his celebrated paper—it was edited and submitted posthumously by Richard Price in 1763. The paper also contains no statement of Bayes' rule, neither in its discrete nor its continuous form. When Bayes' paper was eventually published, it was largely ignored, just as Arbuthnott's paper before. If it had not been for Pierre Simon Laplace (1774), who stated Bayes' rule without reference to Bayes, nobody today would visit Bayes' burial site in Bunhill Fields, in the heart of the city of London. Little is known about Bayes, not even his year of birth, and the only portrait of him shown on hundreds of websites likely displays someone else (Fienberg, 2006). Stigler (1983) estimated a posterior probability of 3 to 1 that not Bayes, but Nicolas Saunderson, the Cambridge Lucasian Professor of Mathematics, discovered what is known today as Bayes' rule: *posterior odds = likelihood ratio × prior odds*, where the likelihood ratio $p(D|H_1)/p(D|H_2)$ is also known as the *Bayes factor*.

Three Interpretations of Probability

Just as the null ritual has replaced three interpretations of level of significance with one, the currently dominant version of Bayesianism does the same with Bayesian pluralism, promoting a universal subjective interpretation instead. But when the mathematical theory of probability was developed in the mid-17th century, there were three interpretations stemming from different applications (Daston, 1988). Probability was

(a) a *relative frequency* in the long run, such as in mortality tables used for calculating insurance premiums;
(b) a *propensity*, that is, the physical design of an object, such as that of a dice or billiard table (which Bayes used as an example); or
(c) a *reasonable degree of subjective belief*, such as in the attempts of courts to quantify the reliability of witness testimony.

In Bayes' essay, his notion of probability is ambiguous and can be read in all three ways. This ambiguity, however, is typical for his time in which the classical theory of probability reigned—from the mid-17th to the early 19th century. The three interpretations were not clearly separated because the classical theory assumed that the beliefs of educated people mirrored objective frequencies (Daston, 1988). In Laplace's famous phrase, probability theory is "only common sense reduced to a calculus" (Laplace 1814/1951: 196). By 1840, however, the classical theory had lost its momentum, mathematicians dissociated their theory from degrees of belief, and the frequency interpretation of probabilities predominated (Gigerenzer et al., 1989). Only a century later—through the writings of the Italian statistician Bruno de Finetti, the British philosopher Frank Ramsey, and the American statistician Leonard Jimmie Savage—did subjective probability experience a renaissance.

A cornerstone of that resurrection is Edwards, Lindman, and Savage's (1963) *Psychological Review* article. In it, they promoted Bayesian statistics instead of what they called classical statistics. For half a century, however, the article had virtually no impact on research method. At that time, the null ritual was already institutionalized, and delusions about the *p* value were in place, including the belief that the *p* value specifies the Bayesian posterior probability that the alternative hypothesis is correct—$p(D|H)$ as surrogate for $p(H|D)$ (Haller & Krauss, 2002).

Universal Bayes

If probability is thought of as a relative frequency in the long run, it immediately becomes clear that Bayes' rule has a limited range of applications. The same holds for propensity. The economist Frank Knight (1921) used the term *risk* for these two situations (i.e., probabilities that can be reliably measured in terms of frequency or propensity) as opposed to *uncertainty*. By contrast, the subjective interpretation has no limits. Subjective probability can be applied to situations of uncertainty and to singular events, such as the probability that Elvis is still alive. Whether that makes sense is another question. Savage (1954/1972: 27) himself was well aware of the problem: Many "are convinced that such statements about probability to a person mean precisely nothing, or at any rate that they mean nothing precisely." Unlike many of his followers, Savage carefully limited Bayesian decision theory to "small worlds" in which all alternatives, consequences, and probabilities are known. And he warned that it would be "utterly ridiculous" to apply Bayesian theory outside a well-defined world—for him, "to plan a picnic" was already outside because the planners cannot know all consequences in advance (Savage, 1954/1972: 16).

This modesty is foreign to a new generation of Bayesians who profess to follow Savage. Instead, they argue that Bayesianism is the only game in town. Earman (1992: 2) saw in this new dogmatism "the imperialistic ambitions of Bayesianism." We will use the term *Universal Bayes* for the view that all uncertainties can or should be represented by subjective probabilities, which explicitly rejects Knight's (1921) distinction between risk and uncertainty. In Lindley's (1983: 10-11) words,

> the Bayesian paradigm concerns uncertainty. . . . It applies to statistical, repetitive situations. . . . But it is also applied to unique situations. We are uncertain about the inflation rate next year, the world's oil reserves, or the possibility of nuclear accidents. All these can be handled by subjective probability.

As a consequence, Universal Bayes ignores the study of genuine tools for uncertainty. These include *exploratory data analysis* (Tukey, 1977), *classification-and-regression trees* (Breiman, Friedman, Olshen, & Stone, 1993), and *fast-and-frugal heuristics* (Gigerenzer, Hertwig, & Pachur, 2011).

Risk Versus Uncertainty

What Universal Bayesians do not seem to realize is that Bayesian theory can be optimal in a world of risk but is of uncertain value in an uncertain world, that is, when not all information is known or can be known or when probabilities have to be estimated from small, unreliable samples. The same point can be made for frequentists. Indeed, under uncertainty, humans rely on simple tools, including the fast-and-frugal heuristics mentioned above, rather than trying to revise subjective probability distributions. And under uncertainty, simple heuristics are actually a better option, given that they can yield more accurate inferences than can complex computational algorithms, including Bayesian ones (Gigerenzer et al., 2011). One can use computer simulations and formal analysis, grounded, for instance, in the *bias-variance dilemma* (Geman, Bienenstock, & Doursat, 1992), to specify *when* less complexity is more—such as when a structural equation model with fewer parameters makes better predictions

than one with more parameters (Gigerenzer & Brighton, 2009). One can also use plain common sense to see that complex optimization algorithms are unreliable in an uncertain world. Take the financial crisis of 2008. The probability models used by banks, regulatory agencies, and rating firms were not the solution to the problem but were in fact part of the problem. They missed the crisis, yet created illusory certainty. As business magnate and billionaire George Soros (2009: 6) bluntly stated, "rational expectations theory is no longer taken seriously outside academic circles." But in our experience, Universal Bayesians tend to take no notice of these results and simply assert that all problems could be best handled by subjective probability, if only the right priors are found. Weick (1996) made the more general point that if scholars or organizations are fixated on "heavy tools," they may become inflexible and fall behind.

Automatic Bayes

As with the null ritual, the universal claim for Bayes' rule tends to go together with its automatic use. In the words of Lindley (1983: 2),

> the Bayesian paradigm provides rules of procedure to be followed. I like to think of it as providing a *recipe*: a set of rules for attaining the final product. The recipe goes like this. What is uncertain and of interest to you? Call it θ. What do you know? Call it D, specific† to the problem, and H, general. Then calculate $p(\theta|D, H)$. How? Using the rules of probability, nothing more, nothing less.

One version of Automatic Bayes is the mechanical interpretation of Bayes factors, using Jeffreys' (1961) scale or similar interpretation aids independent of context (e.g., $B_{10} > 100$ = "decisive evidence against H_0"). Ironically, some of the most-cited propagators of such scales (Kass & Raftery, 1995) point out that this should not be done. Others do not seem to mind mechanical inference (see epigram above).

A second version of Automatic Bayes can be found in the *heuristics-and-biases research program*—a program that is widely taught in business education courses. One of its conclusions is that the mind "is not Bayesian at all" (Kahneman & Tversky, 1972: 450). Instead, people are said to ignore base rates, which is called the *base rate fallacy* and attributed to cognitive limitations. According to these authors, all one has to do to find the correct answer to a textbook problem is to insert the numbers in the problem into Bayes' rule—the content of the problem and content-related assumptions are immaterial. The consequence is a "schizophrenic" split between two standards of rationality: If experimental participants failed to use Bayes' rule to make an inference from a sample, this was considered irrational. But when the researchers themselves made an inference about whether their participants were Bayesians, they did not use Bayes' rule either. Instead, they went through the null ritual, relying only on the p value. In doing so, they themselves committed the base rate fallacy. What is more, subsequent studies revealed that this form of Automatic Bayes mistakes people's good statistical intuitions for cognitive illusions. For instance, participants turned out to be sensitive to relevant assumptions, such as random drawing, whereas researchers who used Automatic Bayes were not (e.g., Gigerenzer, Hell, & Blank, 1988). Moreover, there are multiple alternative Bayesian solutions and Neyman-Pearson solutions to the textbook problems

used in this research and not just one, as asserted by Tversky, Kahneman, and their followers (Gigerenzer & Murray, 1987).

In short, an automatic use of Bayes' rule is a dangerously beautiful idol. But even for a devoted Bayesian, it is not a reality: Like frequentism, Bayesianism does not exist in the singular. According to the mathematician I. J. Good (1971), there are 46,656 varieties of Bayesianism. These differ in beliefs, such as that physical probabilities (propensities or relative frequencies) (a) exist, (b) do not exist, or (c) should be used as if they exist, without philosophical commitment.

Toward a Statistical Toolbox

In our view, the alternative to Universal and Automatic Bayes is to think of Bayesian statistics as forming part of a larger toolbox. In the toolbox view, Bayes' rule has its value but, like any other tool, does not work for all problems.

One of the most useful und least controversial uses of Bayes' rule is in medical diagnosis. As long as the prior odds are based on relative frequencies from epidemiological studies, Bayes' rule can provide useful estimates for the probability of disease given a positive test. Yet most physicians do not know how to use it and are confused by conditional probabilities. The best method for helping physicians learn to think the Bayesian way is to teach them how to translate conditional probabilities into *natural frequencies* (Gigerenzer, 2011). Consider a disease (= hypothesis) H and a positive test (= data) D, with $p(H) = .01$, $p(D|H) = .80$, and $p(D|\neg H) = .04$. Few doctors can infer the posterior probability $p(H|D)$ from these probabilities. To represent them in natural frequencies, one starts with, say, 1,000 patients. We expect that 10 have the disease, 8 of which test positive, and among the 992 without the disease, we expect another ≈40 to test positive. Now it is easy to see that 8 of the 48 who test positive actually have the disease, which translates into $p(H|D) = .17$. Today, major medical organizations, such as the Cochrane Collaboration and the Medicine and Healthcare Products Regulatory Agency of the United Kingdom, recommend using natural frequencies.

However, even frequency-based probabilities are not universally accepted. Take the credibility of eyewitness testimony in court, which was the historical origin of the subjective interpretation of probability in the 17th century. Not every eyewitness is trustworthy, and so one tried to quantify the probability that a witness is telling the truth (Daston, 1988). Despite this historical relation between probability and the law, to the present day, legal professionals often do not understand statistical evidence—for instance, DNA fingerprinting and wife battering in the O. J. Simpson case (Gigerenzer, 2002: chaps. 8-10). Most important, courts in the United States and Great Britain have often resisted the introduction of Bayes' rule (Koehler, 1992). Prior probabilities other than zero are considered inconsistent with the presumption of innocence. And for many legal systems, statistical justice is a contradiction in terms because it replaces the individual with averages, the individual case with a generality, and, most important, the exercise of judgment with the application of automatic rules (Gigerenzer et al., 1989).

In the social sciences, objections to the use of Bayes' rule are that frequency-based prior probabilities do not exist, that the set of hypotheses needed for the prior probability distribution is not known, and that researchers' introspection does not confirm the calculation of probabilities. Few researchers report that they consider the set of all plausible hypotheses and attach probabilities to these, summing up to 1. Fisher thought that prior probabilities were

meaningful only when they can be empirically estimated, which is rare for scientific hypotheses. In his opinion, Bayes "deserves honourable remembrance" for realizing the limits of his approach and "perhaps, for the same reason, [for deciding] to withhold his entire treatise from publication" (Fisher, 1935/1971: 6). Moreover, as has been pointed out by Herbert A. Simon, scientists do not assign cardinal numbers to subjective probabilities in an uncertain world. Instead, they rely on fast-and-frugal heuristics that are robust in the many situations in which the assumptions for Bayes' rule are not met (e.g., Simon, 1979; Gigerenzer et al., 2011). The null ritual provides, in a negative sense, testimony to this: It can be seen as the mindless institutionalization of a single, simple rule.

In sum, unlike those who worship the idol of a universal method, we argue that Bayes' rule is useful as part of a statistical toolbox, for instance, when priors can be reliably estimated, as in medical diagnosis based on epidemiological studies. Neyman and Pearson's decision theory, like Wald's sequential decision theory, is appropriate for repeated random drawing situations in quality control, such as in Newton's work as the master of the Royal Mint. Fisher's null hypothesis testing is another tool, relevant for situations in which one does not understand what is happening, such as when precise alternative hypotheses cannot be specified. This statistical toolbox contains not only techniques of inference but—of equal importance—descriptive statistics, exploratory data analysis, and formal modeling techniques as well. The only items that do not belong in the toolbox are false idols.

How Statistics Change Research—Surrogate Science

As a young man, Gottfried Wilhelm Leibniz (1677/1951) had a dream: to discover the calculus that could map all ideas into symbols. Such a universal calculus would put an end to all scholarly bickering—if a dispute arose, the contending parties could settle it peacefully by saying, "Let's sit down and calculate." Leibniz optimistically guessed that a few skilled persons might be able to complete the whole thing in five years. Yet neither he nor anyone else has succeeded.

Leibniz's dream is nonetheless still alive in the social sciences. Because the object of the dream has not been found, surrogates serve in its place. In some fields, it is the routine use of significance; in others, Bayesian statistics. We use the term *surrogate science* in a more general sense, indicating the attempt to *infer* the quality of research using a single number or benchmark. The introduction of surrogates shifts researchers' goal away from doing innovative science and redirects their efforts toward meeting the surrogate goal. Let us give a few examples.

Statistical Inference as Surrogate for Replication

A significant p value does not specify the probability that the same result can be reproduced in another study. That can easily be seen from the fact that a p value, that is, the probability $p(D|H_0)$, is not the same as $p(D)$, the probability of the data. Nevertheless, in the minds of some, the p value has become a surrogate for replicability: If $p = .01$, then the probability is .99 that another significant result will be obtained in a replication. Studies reported that this delusion was believed to be true by about half of psychology department teachers (Haller & Krauss, 2002; Oakes, 1986). Textbooks have also spread this delusion, known as the *replication fallacy* (Gigerenzer, 2004).

Consistent with its surrogate function, journal articles report few replications in comparison to the number of *p* values. An analysis of 835 empirical articles sampled from the *Journal of Marketing*, the *Journal of Marketing Research*, and the *Journal of Consumer Research* found no single straight replications and in only 2.4% of them replications with extensions (Hubbard & Armstrong, 1994). An analysis of 1,046 research articles in the *Journal of Consulting and Clinical Psychology, Journal of Counseling Psychology*, and *Personnel and Guidance Journal* revealed that 86% used statistical tests of significance, 94% of these rejected the null at $p < .05$ or less, but fewer than 1% of them were replications (Bozarth & Roberts, 1972).

Traditional *p* values are not the only surrogates for real replications: For some time, the editors of *Psychological Science* even recommended that researchers routinely report *p-rep*, a mere transformation of the *p* value, in order to estimate the "probability of replicating an effect" ("Information for Contributors," 2005: i). Advocates of Bayesian statistics rightly criticized this delusion but called instead for automatic Bayes factors: "It should be the routine business of authors contributing to *Psychological Science* or any other Journal of scientific psychology to report Bayes Factors" (Iverson, Lee, Zhang, & Wagenmakers, 2009: 201).

Inferential statistics have become surrogates for real replication. The consequence is a flood of irreproducible results. Analyses of replications in management, finance, and advertising journals showed that 40% to 60% of these contradicted the results of the original studies (see Hubbard & Armstrong, 1994). Scientists at Amgen were unable to replicate 47 out of 53 major oncological findings for potential drug targets, nor were researchers at Bayer able to replicate 43 out of 67 oncological and cardiovascular findings (see Ioannidis et al., 2014). Psychologists have encountered similar surprises when trying to replicate "significant" findings.

Hypotheses Finding Is Presented as Hypotheses Testing

If the mean number of *p* values in an article is 99 (see introduction), warning bells should ring. For one, the authors may have disguised hypothesis finding as hypothesis testing, a practice known as *fishing expeditions*. Unfortunately, to the present day, many researchers first look at the data for patterns, check for significance, and then present the result as if it were a hypothesis test (Kerr, 1998). SPSS and other user-friendly software packages that automatically run tests facilitate this form of scientific misconduct: A hypothesis should not be tested with the same data from which it was derived. Finding new patterns is important, but *p* values or confidence intervals should not be provided for these. The same argument holds for subgroup tests after the fact (e.g., controlling for gender or gender interactions; P. Good & Harding, 2003) as well as for Bayesian methods that use the data to estimate the prior after the data have been seen. A similarly bad practice, common in management, education, and sociology, is to routinely fit regressions and other statistical models to data, report R^2 and significance, and stop there—without subjecting the results to actual tests via cross-validation or other model selection procedures.

We propose a simple rule of thumb—let's call it the *f-index*—to infer the extent *f* of fishing expeditions disguised as hypothesis testing: $f = n_t/n_h$, where n_h is the number of hypotheses stated in an article and n_t is the number of statistical tests, such as the number of reported *p* values or confidence intervals. If $f = 1$, there is no sign of a fishing expedition; if $f > 1$, there

is. The *f*-index is a conservative estimate because nonsignificant *p* values or confidence intervals including the zero, looked up after the fact, may not be reported. The *f*-index is also relevant for Bayesian statistics when researchers use software packages that automatically churn out Bayes factors.

In sum, routine statistical inference has become a surrogate for both hypothesis finding and replication. The surrogate goal is to obtain a significant *p* value or other test statistic, even when it is out of place, as in the case of hypothesis finding.

Quantity as Surrogate for Quality

Surrogate science does not end with statistical tests. Research assessment exercises tend to create surrogates as well. Citation counts, impact factors, and *h*-indices are also "inferential statistics" that administrators and search committees may (ab)use to infer the quality of research. Just as statistical software (e.g., SPSS, STATA) facilitates computing *p* values and Bayes factors, digital media (e.g., Google Scholar, Web of Knowledge) make surrogates easily accessible.

The evident danger is that hiring committees and advisory boards study these surrogate numbers rather than the papers written by job candidates and faculty members. If being cited meant being read, citation statistics might well be a useful criterion. Yet a study estimated that of the articles cited, only 20% had actually been read (Simkin & Roychowdhury, 2003). With citation as a surrogate for quality, some truly original work may go unheeded. For instance, the most important publication in 20th-century biology, Watson and Crick's paper on the double helix, was rarely cited in the first 10 years after its publication (Olby, 2003). Innovative ideas take time to be appreciated.

An even greater danger is that surrogates transform science by warping researchers' goals. If a university demands publication of *X* journal articles for promotion, this number provides an incentive for researchers to dissect a coherent paper into small pieces for several journals. These pieces are aptly called *just publishable units*. Peter Higgs, the 2013 Nobel Prize winner in physics, once said in an interview, "Today I wouldn't get an academic job. It's as simple as that. I don't think I would be regarded as productive enough" (Aitkenhead, 2013). He added that because he was not churning out papers as expected at Edinburgh University, he had become "an embarrassment to the department when they did research assessment exercises" (Aitkenhead, 2013).

Conclusion: Leibniz's Dream or Bayes' Nightmare?

Surrogate science, from the mindless calculation of *p* values or Bayes factors to citation counts, is not entirely worthless. It fuels a steady stream of work of average quality and keeps researchers busy producing more of the same. But it makes it harder for scientists to be innovative, risk taking, and imaginative.

Even worse, by transforming researchers' goals, surrogates also encourage cheating and incomplete or dishonest reporting. In a survey of over 2,000 academic psychologists at major U.S. universities, almost half admitted to having selectively reported studies that "worked." Equally alarming, more than half admitted having decided whether to collect more data *after* having examined whether the results were significant (John, Loewenstein, & Prelec, 2012). To illustrate the consequences, a researcher who has already collected 10 observations in each of

two conditions and now conducts a *t* test after each new observation (per condition) can obtain a false-positive result ($p \leq .05$) 22% of the time (Simmons, Nelson, & Simonsohn, 2011). Selection of correlations after the fact appears to be common practice in social neurosciences, resulting in "voodoo correlations" (Vul, Harris, Winkielman, & Pashler, 2009: 274)—correlations between brain activity and psychological measures that are higher than they can possibly be, given the limited reliability of the measures. These practices are not outright fraud but borderline cheating. Yet they are likely a greater hindrance to progress than fraud, which is comparatively rare. Their product is an impressive flood of "significant" but irreproducible results.

Would a Bayesian revolution lead to a better world? The answer depends on what the revolution might be. As with the methods of Fisher and Neyman and Pearson, the real challenge in our view is to prevent the same surrogates from taking over once again, such as when replacing routine significance tests with routine interpretations of Bayes factors. Otherwise, Leibniz's beautiful dream of a universal calculus could easily turn into "Bayes' nightmare." Editorials such as that by Altmann (2007: 6), cited in the epigram to this article, could be easily rewritten in favor of the old idol:

> *Cognition* will require its authors to adhere to the convention that ~~an effect~~ evidence *described* as "~~statistically significant~~" "very strong" must ~~have~~ come with ~~a *p* value below .05~~ Bayes factors 30–100 for H_0 and 1/100–1/30 for H_1 (for better or for worse, this is the current convention).

Much ink has been spilled by Bayesians in criticizing frequentists, and vice versa. But the true enemy lies within each of the fighting parties. It is the idol of a universal method of scientific inference.

Notes

1. This analysis includes all empirical articles, be they qualitative or quantitative.

2. Mechanical reporting of *p* values is fortunately not everywhere. Back in 1988, the guidelines of the International Committee of Medical Journal Editors instructed, "Avoid sole reliance on statistical hypothesis testing, such as the use of *p* values" (p. 402). And in 1998, Rothman, the editor of *Epidemiology*, stated, "When writing for *Epidemiology*, you can also enhance your prospects if you omit tests of statistical significance. . . . Every worthwhile journal will accept papers that omit them entirely. In *Epidemiology*, we do not publish them at all" (p. 334).

3. For instance, in the 19th century, significance tests were occasionally used in astronomy—but typically for rejecting observations (outliers), not hypotheses (Gigerenzer, Swijtink, Porter, Daston, Beatty, & Krüger, 1989). This illustrates that statistical tests can be used for two different purposes. One is to trust in a hypothesis (such as in a normal distribution of observational errors around the true position of a star) and reject observations that deviate too far from it, possibly caused by a distracted or tired observer. The second is the reverse: to trust in the observations and reject hypotheses that deviate too far from them. The latter has become virtually the only use of statistical inference in the social sciences. Unlike astronomers, social scientists approach the question of outliers by applying subjective judgment, not statistical inference.

References

Aitkenhead, D. 2013, December 6. Peter Higgs: I wouldn't be productive enough for today's academic system. *The Guardian*. Retrieved from http://www.theguardian.com/science/2013/dec/06/peter-higgs-boson-academic-system

Altmann, G. 2007. Editorial journal policies and procedures. *Cognition*, 102: 1-6.

American Psychological Association. 1974. *Publication manual of the American Psychological Association*, 2nd ed. Baltimore, MD: Garamond/Pridemark.

Arbuthnott, J. 1710. An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society*, 27: 186-190.

Bacon, F. 1902. *Novum organum*. New York: P. F. Collier & Son. (Original work published 1620)

Boring, E. G. 1963. *History, psychology, and sciences: Selected papers*. New York: Wiley.

Bozarth, J. D., & Roberts, R. R., Jr. 1972. Signifying significant significance. *American Psychologist*, 27: 774-775.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. 1993. *Classification and regression trees*. New York: Chapman & Hall.

Danziger, K. 1990. *Constructing the subject: Historical origins of psychological research*. Cambridge, UK: Cambridge University Press.

Daston, L. 1988. *Classical probability in the Enlightenment*. Princeton, NJ: Princeton University Press.

Daston, L. 2005. Scientific error and the ethos of belief. *Social Research*, 72: 1-28.

Earman, J. 1992. *Bayes or bust? A critical examination of Bayesian confirmation theory*. Cambridge, MA: MIT Press.

Edwards, W., Lindman, H., & Savage, L. J. 1963. Bayesian statistical inference for psychological research. *Psychological Review*, 70: 193-242.

Fienberg, S. E. 2006. When did Bayesian inference become "Bayesian"? *Bayesian Analysis*, 1: 1-40.

Fisher, R. A. 1955. Statistical methods and scientific induction. *Journal of the Royal Statistical Society*, 17 (Series B): 69-77.

Fisher, R. A. 1956. *Statistical methods and scientific inference*. Edinburgh, UK: Oliver & Boyd.

Fisher, R. A. 1971. *The design of experiments*, reprint of the 8th ed. New York: Hafner. (Original work published 1935)

Franco, Z. E., Blau, K., & Zimbardo, P. G. 2011. Heroism: A conceptual analysis and differentiation between heroic action and altruism. *Review of General Psychology*, 15: 99-113.

Geman, S., Bienenstock, E., & Doursat, R. 1992. Neural networks and the bias/variance dilemma. *Neural Computation*, 4: 1-58.

Gerrig, R. J., & Zimbardo, P. G. 2002. *Psychology and life*, 16th ed. Boston: Allyn & Bacon.

Gigerenzer, G. 1987. Probabilistic thinking and the fight against subjectivity. In L. Krüger, G. Gigerenzer, & M. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences*, 11-33. Cambridge, MA: MIT Press.

Gigerenzer, G. 2002. *Calculated risks: How to know when numbers deceive you*. New York: Simon & Schuster.

Gigerenzer, G. 2004. Mindless statistics. *Journal of Socio-Economics*, 33: 587-606.

Gigerenzer, G. 2011. What are natural frequencies? *British Medical Journal*, 343: d6386.

Gigerenzer, G., & Brighton, H. 2009. Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1: 107-143.

Gigerenzer, G., Hell, W., & Blank, H. 1988. Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance*, 14: 513-525.

Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). 2011. *Heuristics: The foundations of adaptive behavior*. Oxford, UK: Oxford University Press.

Gigerenzer, G., & Murray, D. J. 1987. *Cognition as intuitive statistics*. Hillsdale, NJ: Lawrence Erlbaum.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. 1989. *The empire of chance: How probability changed science and everyday life*. Cambridge, UK: Cambridge University Press.

Good, I. J. 1971. 46656 varieties of Bayesianism. *American Statistician*, 25: 62-63.

Good, P. I., & Harding, J. W. 2003. *Common errors in statistics (and how to avoid them)*. New York: Wiley.

Guilford, J. P. 1942. *Fundamental statistics in psychology and education*. New York: McGraw-Hill.

Hacking, I. 1987. Prussian numbers 1860-1882. In L. Krüger, L. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history*: 377-394. Cambridge, MA: MIT Press.

Haller, H., & Krauss, S. 2002. Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research—Online*, 7: 1-20.

Hubbard, R., & Armstrong, J. S. 1994. Replications and extensions in marketing: Rarely published but quite contrary. *International Journal of Research in Marketing*, 11: 233-248.

Information for contributors. 2005. *Psychological Science*, 16: i.

International Committee of Medical Journal Editors. 1988. Uniform requirements for manuscripts submitted to biomedical journals. *British Medical Journal*, 296: 401-405.

Ioannidis, J. P. A., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., . . .Tibshirani, R. 2014. Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet*, 383: 166-175.

Iverson, G. J., Lee, M. D., Zhang, S., & Wagenmakers, E. J. 2009. Prep: An agony in five Fits. *Journal of Mathematical Psychology*, 53: 195-202.

Jeffreys, H. 1961. *Theory of probability*. Oxford, UK: Oxford University Press.

John, L., Loewenstein, G. F., & Prelec, D. 2012. Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, 23: 524-532.

Kahneman, D., & Tversky, A. 1972. Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3: 430-454.

Kass, R. E., & Raftery, A. E. 1995. Bayes factors. *Journal of the American Statistical Association*, 90: 773-795.

Kerr, N. L. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2: 196-217.

Knight, F. H. 1921. *Risk, uncertainty, and profit*. New York: Houghton Mifflin.

Koehler, J. J. 1992. Probabilities in the courtroom: An evaluation of the objections and policies. In D. K. Kagehiro & W. S. Laufer (Eds.), *Handbook of psychology and law*: 167-184. New York: Springer-Verlag.

Krüger, L., Daston, L. J., & Heidelberger, M. (Eds.). 1987. *The probabilistic revolution: Vol. 1. Ideas in history*. Cambridge, MA: MIT Press.

Krüger, L., Gigerenzer, G., & Morgan, M. S. (Eds.). 1987. *The probabilistic revolution: Vol. 2. Ideas in the sciences*. Cambridge, MA: MIT Press.

Laplace P. S. 1774. Mémoire sur la probabilité des causes par les évènemens. *Mémoires de mathématique et de physique, presentés a l'Académie Royale des Sciences, par divers savans, & lû dans ses assemblées*, 6: 621-656. (English translation in Stigler, S. M. 1986. Laplace's 1774 memoir on inverse probability. *Statistical Science*, 1: 359-378.)

Laplace, P. S. 1951. *A philosophical essay on probabilities*. New York: Dover. (Original work published 1814)

Leibniz, G. W. 1951. Toward a universal characteristic. In P. P. Wiener (Ed.), *Selections*: 17-25. New York: Scribner's. (Original work published 1677)

Lindley, D. V. 1975. The future of statistics: A Bayesian 21st century. *Advances in Applied Probability (Supplement)*, 7: 106-115.

Lindley, D. V. 1983. Theory and practice of Bayesian statistics. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32: 1-11.

Lindley, D. V. 1986. Comment. *American Statistician*, 40: 6-7.

Luce, R. D. 1988. The tools-to-theory hypothesis. Review of G. Gigerenzer & D. J. Murray, "Cognition as Intuitive Statistics." *Contemporary Psychology*, 33: 582-583.

Madden, C. S., Easley, R. W., & Dunn, M. G. 1995. How journal editors view replication research. *Journal of Advertising*, 24: 77-79.

Maslow, A. H. 1966. *The psychology of science*. New York: Harper & Row.

Melton, A. W. 1962. Editorial. *Journal of Experimental Psychology*, 64: 553-557.

Newton, I. 1952. *Opticks. Or a treatise of the reflections, refractions, inflections, and colours of light*. New York: Dover. (Original work published 1704)

Neyman, J. 1957. Inductive behavior as a basic concept of philosophy of science. *International Statistical Review*, 25: 7-22.

Nunnally, J. C. 1975. *Introduction to statistics for psychology and education*. New York: McGraw-Hill.

Oakes, M. 1986. *Statistical inference: A commentary for the social and behavioral sciences*. Chichester, UK: Wiley.

Olby, R. 2003. Quiet debut for the double helix. *Nature*, 276: 565-570.

Rothman, K. J. 1998. Writing for *Epidemiology. Epidemiology*, 9: 333-337.

Savage, L. J. 1972. *The foundations of statistics*. New York: Dover (Original work published 1954)

Simkin, M. V., & Roychowdhury, V. P. 2003. Read before you cite! *Complex Systems*, 14: 269-274.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 20: 1-8.

Simon, H. A. 1979. Rational-decision making in business organizations (Nobel Memorial Lecture, 8 December, 1978). *American Economic Review*, 69: 493-513.

Simon, H. A. 1992. What is an "explanation" of behavior? *Psychological Science*, 3: 150-161.

Soros, G. 2009. *The crash of 2008 and what it means: The new paradigm for financial markets*. New York: Public Affairs.

Stegmüller, W. 1973. *Jenseits von Popper und Carnap* [Beyond Popper and Carnap]. Berlin, Germany: Springer.

Stevens, S. S. 1960. The predicament in design and significance. *Contemporary Psychology*, 9: 273-276.

Stigler, S. M. 1983. Who discovered Bayes's theorem? *American Statistician*, 37: 290-296.

Stigler, S. M. 1999. *Statistics on the table: The history of statistical concepts and methods*. Cambridge, MA: Harvard University Press.

Style guide for authors. 2011. *Academy of Management Journal*, 54: 1081-1084.

Tressoldi, P. E., Giofré, D., Sella, F., & Cumming, G. 2013. High impact = high statistical standards? Not necessarily so. *PLOS ONE*, 8: 1-7.

Trotta, R. 2007. Applications of Bayesian model selection to cosmological parameters. *Monthly Notices of the Royal Astronomical Society*, 378: 72-82.

Tukey, J. W. 1977. *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, 4: 274-290.

Weick, K. E. 1996. Drop your tools: An allegory for organizational studies. *Administrative Science Quarterly*, 41: 301-313.

Ziliak, S. T., & McCloskey, D. N. 2004. Size matters: The standard error of regressions in the *American Economic Review*. *Journal of Socio-Economics*, 33: 527-546.