ELSEVIER

The Journal of Socio-Economics 33 (2004) 587–606

The Journal of Socio-Economics

www.elsevier.com/locate/econbase

Mindless statistics

Gerd Gigerenzer*

Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

Abstract

Statistical rituals largely eliminate statistical thinking in the social sciences. Rituals are indispensable for identification with social groups, but they should be the subject rather than the procedure of science. What I call the "null ritual" consists of three steps: (1) set up a statistical null hypothesis, but do not specify your own hypothesis nor any alternative hypothesis, (2) use the 5% significance level for rejecting the null and accepting your hypothesis, and (3) always perform this procedure. I report evidence of the resulting collective confusion and fears about sanctions on the part of students and teachers, researchers and editors, as well as textbook writers.

Keywords: Rituals; Collective illusions; Statistical significance; Editors; Textbooks

> . . . 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)

I once visited a distinguished statistical textbook author, whose book went through many editions, and whose name does not matter. His textbook represents the relative best in the social sciences. He was not a statistician; otherwise, his text would likely not have been used in a psychology class. In an earlier edition, he had included a chapter on Bayesian statistics, and also mentioned (albeit in only one sentence) that there was a development in statistical theory from R.A. Fisher to Jerzy Neyman and Egon S. Pearson. To mention the existence of alternative methods and the names associated with them is virtually unheard

* Tel.: +49 30 82406 460; fax: +49 30 82406 394.
E-mail address: gigerenzer@mpib-berlin.mpg.de.

1053-5357/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.socec.2004.09.033

of in psychology. I asked the author why he removed the chapter on Bayes as well as the innocent sentence from all subsequent editions. "What made you present statistics as if it had only a single hammer, rather than a toolbox? Why did you mix Fisher's and Neyman–Pearson's theories into an inconsistent hybrid that every decent statistician would reject?"

To his credit, I should say that the author did not attempt to deny that he had produced the illusion that there is only one tool. But he let me know who was to blame for this. There were three culprits: his fellow researchers, the university administration, and his publisher. Most researchers, he argued, are not really interested in statistical thinking, but only in how to get their papers published. The administration at his university promoted researchers according to the number of their publications, which reinforced the researchers' attitude. And he passed on the responsibility to his publisher, who demanded a single-recipe cookbook. No controversies, please. His publisher had forced him to take out the chapter on Bayes as well as the sentence that named alternative theories, he explained. At the end of our conversation, I asked him what kind of statistical theory he himself believed in. "Deep in my heart," he confessed, "I am a Bayesian."

If the author was telling me the truth, he had sold his heart for multiple editions of a famous book whose message he did not believe in. He had sacrificed his intellectual integrity for success. Ten thousands of students have read his text, believing that it reveals the method of science. Dozens of less informed textbook writers copied from his text, churning out a flood of offspring textbooks, and not noticing the mess.

1. The null ritual

Textbooks and curricula in psychology almost never teach the statistical toolbox, which contains tools such as descriptive statistics, Tukey's exploratory methods, Bayesian statistics, Neyman–Pearson decision theory and Wald's sequential analysis. Knowing the contents of a toolbox, of course, requires statistical thinking, that is, the art of choosing a proper tool for a given problem. Instead, one single procedure that I call the "null ritual" tends to be featured in texts and practiced by researchers. Its essence can be summarized in a few lines:

The null ritual:

1. Set up a statistical null hypothesis of "no mean difference" or "zero correlation." Don't specify the predictions of your research hypothesis or of any alternative substantive hypotheses.
2. Use 5% as a convention for rejecting the null. If significant, accept your research hypothesis. Report the result as $p < 0.05$, $p < 0.01$, or $p < 0.001$ (whichever comes next to the obtained p-value).
3. Always perform this procedure.

The null ritual has sophisticated aspects I will not cover here, such as alpha adjustment and ANOVA procedures. But these do not change its essence. Often, textbooks also teach concepts alien to the ritual, such as statistical power and effect sizes, but these additions tend

to disappear when examples are given. They just don't fit. More recently, the ritual has been labeled *null hypothesis significance testing*, for short, *NHST* or sometimes *NHSTP* (with *P* for "procedure"). It became institutionalized in curricula, editorials, and professional associations in psychology in the mid-1950s (Gigerenzer, 1987, 1993). The 16th edition of a highly influential textbook, Gerrig and Zimbardo's *Psychology and Life* (2002), portrays the null ritual as statistics per se and calls it the "backbone of psychological research" (p. 46). Its mechanical nature is sometimes presented like the rules of grammar. For instance, the 1974 *Publication Manual of the American Psychological Association* told authors what to capitalize, when to use a semicolon, and how to abbreviate states and territories. It also told authors how to interpret *p*-values: "Caution: Do not infer trends from data that fail by a small margin to meet the usual levels of significance. Such results are best interpreted as caused by chance and are best reported as such. Treat the result section like an income tax return. Take what's coming to you, but no more" (p. 19; this passage was deleted in the 3rd ed., 1983). Judgment is not invited. This reminds me of a maxim regarding the critical ratio, the predecessor of the significance level: "A critical ratio of three, or no Ph.D."

Anonymity is essential. The ritual is virtually always presented without names, as statistics per se. If names such as Fisher or Pearson are mentioned in textbooks in psychology, they are usually done so in connection with a minor detail, such as to thank E.S. Pearson for the permission to reprint a table. The major ideas are presented anonymously, as if they were given truths. Which text written for psychologists points out that null hypothesis testing was Fisher's idea? And that Neyman and Pearson argued against null hypothesis testing? If names of statisticians surface, the reader is typically told that they are all of one mind. For instance, in response to a paper of mine (Gigerenzer, 1993), the author of a statistical textbook, S.L. Chow (1998), acknowledged that different methods of statistical inference in fact exist. But a few lines later he fell back into the "it's-all-the-same" fable: "To K. Pearson, R. Fisher, J. Neyman, and E.S. Pearson, NHSTP was what the empirical research was all about" (Chow, 1998, p. xi). Reader beware. Each of these eminent statisticians would have rejected the null ritual as bad statistics.

Fisher is mostly blamed for the null ritual. But toward the end of his life, Fisher (1955, 1956) rejected each of its three steps. First, "null" does not refer to a nil mean difference or zero correlation, but to any hypothesis to be "nullified." A correlation of 0.5, or a reduction of five cigarettes smoked per day, for instance, can be a null hypothesis. Second, as the epigram illustrates, by 1956, Fisher thought that using a routine 5% level of significance indicated lack of statistical sophistication. No respectable researcher would use a constant level. Your chances of finding this quote in a statistical text in psychology is virtually nil. Third, for Fisher, null hypothesis testing was the most primitive type of statistical analyses and should be used only for problems about which we have *no or very little knowledge* (Gigerenzer et al., 1989, chapter 3). He proposed more appropriate methods for other cases. Neyman and Pearson would have also rejected the null ritual, but for different reasons. They rejected null hypothesis testing, and favored competitive testing between two or more statistical hypotheses. In their theory, "hypotheses" is in the plural, enabling researchers to determine the Type-II error (which is not part of the null ritual, and consequently, not germane to NHSTP, as Chow asserts). The confusion between the null ritual and Fisher's theory, and sometimes even Neyman–Pearson theory, is the rule rather than the exception among psychologists.

Psychology seems to be one of the first disciplines where the null ritual became institutionalized as statistics per se, during the 1950s (Rucci and Tweney, 1980; Gigerenzer and Murray, 1987, chapter 1). Subsequently, it spread to many social, medical, and biological sciences, including economics (McCloskey and Ziliak, 1996), sociology (Morrison and Henkel, 1970), and ecology (Anderson et al., 2000).

If psychologists are so smart, why are they so confused? Why is statistics carried out like compulsive hand washing? My answer is that the ritual requires confusion. To acknowledge that there is a statistical toolbox rather than one hammer would mean its end, as would realizing that the null ritual is practiced neither in the natural sciences, nor in statistics proper. Its origin is in the minds of statistical textbook writers in psychology, education, and other social sciences. It was created as an inconsistent hybrid of two competing theories: Fisher's null hypothesis testing and Neyman and Pearson's decision theory.

2. What Fisher and Neyman–Pearson actually proposed

In discussions about the pros and cons of significance testing in the social sciences, it is commonly overlooked (by both sides) that the ritual is not even part of statistics proper. So let us see what Fisher and Neyman–Pearson actually proposed. The logic of Fisher's (1955, 1956) null hypothesis testing can be summarized in three steps:

Fisher's null hypothesis testing:

1. Set up a statistical null hypothesis. The null need not be a nil hypothesis (i.e., zero difference).
2. Report the exact level of significance (e.g., $p = 0.051$ or $p = 0.049$). Do not use a conventional 5% level, and do not talk about accepting or rejecting hypotheses.
3. Use this procedure only if you know very little about the problem at hand.

Fisher's null hypothesis testing is, at each step, unlike the null ritual, but also unlike Neyman–Pearson decision theory. It lacks a specified statistical alternative hypothesis. As a consequence, the concepts of statistical power, Type-II error rates, and theoretical effect sizes have no place in Fisher's framework—one needs a specified alternative for these concepts. The Polish mathematician Jerzy Neyman worked with Egon S. Pearson (the son of Karl Pearson) at University College in London and later, when the tensions between Fisher and himself grew too heated, moved to Berkeley, California. Neyman and Pearson criticized Fisher's null hypothesis testing for several reasons, including that no alternative hypothesis is specified (Gigerenzer et al., 1989, chapter 3). In its simplest version, Neyman–Pearson theory has two hypotheses and a binary decision criterion (Neyman, 1950, 1957).

Neyman–Pearson decision theory:

1. Set up two statistical hypotheses, H_1 and H_2, and decide about α, β, and sample size before the experiment, based on subjective cost-benefit considerations. These define a rejection region for each hypothesis.

2. If the data falls into the rejection region of H_1, accept H_2; otherwise accept H_1. Note that accepting a hypothesis does not mean that you believe in it, but only that you act as if it were true.
3. The usefulness of the procedure is limited among others to situations where you have a disjunction of hypotheses (e.g., either $\mu_1 = 8$ or $\mu_2 = 10$ is true) and where you can make meaningful cost-benefit trade-offs for choosing alpha and beta.

A typical application of Neyman–Pearson testing is in quality control. Imagine a manufacturer of metal plates that are used in medical instruments. She considers a mean diameter of 8 mm (H_1) as optimal and 10 mm (H_2) as dangerous to the patients and hence unacceptable. From past experience, she knows that the random fluctuations of diameters are approximately normally distributed and that the standard deviations do not depend on the mean. This allows her to determine the sampling distributions of the mean for both hypotheses. She considers false alarms, that is, accepting H_2 while H_1 is true, to be the less serious error, and misses of malfunctioning, that is, accepting H_1 while H_2 is true, to be more serious. Misses may cause harm to patients and to the firm's reputation. Therefore, she sets the first error rate small, and the second larger, say $\alpha = 0.1\%$, and $\beta = 10\%$, respectively. She now calculates the required sample size n of plates that must be sampled every day to test the quality of the production (see Cohen, 1988). When she accepts H_2, she acts as if there were a malfunction and stops production, but this does not mean that she believes that H_2 is true. She knows that she must expect a false alarm in 1 out of 10 days in which there is no malfunction (Gigerenzer et al., 1989, chapter 3).

Now it is clear that the null ritual is a hybrid of the two theories. The first step of the ritual, to set up only one statistical hypothesis (the null), stems from Fisher's theory, except that the null always means "chance," such as a zero difference. This first step is inconsistent with Neyman–Pearson theory; it does not specify an alternative statistical hypotheses, α, β, or the sample size. The second step, making a yes–no decision, is consistent with Neyman–Pearson theory, except that the level should not be fixed by convention but by thinking about α, β, and the sample size. Fisher (1955) and many statisticians after him (see Perlman and Wu, 1999), in contrast, argued that unlike in quality control, yes–no decisions have little role in science; rather, scientists should communicate the exact level of significance. The third step of the null ritual is unique in statistical theory. If Fisher and Neyman–Pearson agreed on anything, it was that statistics should never be used mechanically.

Fisher is the best known of the inadvertent "fathers" of the null ritual. His influence has divided psychologists deeply, and interestingly, the rift runs between the great personalities in psychology on the one hand, and a mass of anonymous researchers on the other. You would not have caught Jean Piaget calculating a *t*-test. The seminal contributions by Frederick Bartlett, Wolfgang Köhler, and the Noble laureate I.P. Pavlov did not rely on *p*-values. Stanley S. Stevens, a founder of modern psychophysics, together with Edwin Boring, known as the "dean" of the history of psychology, blamed Fisher for a "meaningless ordeal of pedantic computations" (Stevens, 1960, p. 276). The clinical psychologist Paul Meehl (1978, p. 817) called routine null hypothesis testing "one of the worst things that ever happened in the history of psychology," and the behaviorist B.F. Skinner blamed Fisher and his followers for having "taught statistics in lieu of scientific method" (Skinner, 1972, p. 319). The mathematical psychologist R. Duncan Luce (1988, p. 582) called null hypothesis testing

a "wrongheaded view about what constituted scientific progress" and the Nobel laureate Herbert A. Simon (1992, p. 159) simply stated that for his research, the "familiar tests of statistical significance are inappropriate."

It is telling that few researchers are aware that their own heroes rejected what they practice routinely. Awareness of the origins of the ritual and of its rejection could cause a virulent cognitive dissonance, in addition to dissonance with editors, reviewers, and dear colleagues. Suppression of conflicts and contradicting information is in the very nature of this social ritual.

3. Feelings of guilt

Let me introduce Dr. Publish-Perish. He is the average researcher, a devoted consumer of statistical packages. His superego tells him that he ought to set the level of significance before an experiment is performed. A level of 1% would be impressive, wouldn't it? Yes, but ... He fears that the *p*-value calculated from the data could turn out slightly higher. What if it were 1.1%? Then he would have to report a nonsignificant result. He does not want to take that risk. How about setting the level at a less impressive 5%? But what if the *p*-value turned out to be smaller than 1% or even 0.1%? He would then regret his decision deeply, because he would have to report this result as $p < 0.05$. He does not like that either. So he concludes that the only choice left is to cheat a little and disobey his superego. He waits until he has seen the data, rounds the *p*-value up to the next conventional level, and reports that the result is significant at $p < 0.001$, 0.01, or 0.05, whatever is next. That smells of deception, and his superego leaves him with feelings of guilt. But what should he do when honesty does not pay, and nearly everyone else plays this little cheating game?

Dr. Publish-Perish does not know that his moral dilemma is caused by a mere confusion, introduced by textbook writers who failed to distinguish the three main interpretations of the level of significance.

3.1. Level of significance = mere convention

Fisher wrote three books on statistics. For the social sciences, the most influential of them was the second one, the *Design of Experiments*, first published in 1935. Fisher's definition of a level of significance differed here from his later writings. In the *Design*, Fisher suggested that we think of the level of significance as a *convention*: "It is usual and convenient for experimenters to take 5% as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard" (1935/1951, p. 13). Fisher's assertion that 5% (in some cases, 1%) is a convention to be adopted by all experimenters and in all experiments, while nonsignificant results are to be ignored, became part of the null ritual.

3.2. Level of significance = alpha

In Neyman–Pearson theory, the meaning of a level of significance such as 2% is the following: If H_1 is correct, and the experiment is repeated many times, the experimenter

will wrongly reject H_1 in 2% of the cases. Rejecting H_1 if it is correct is called a Type-I error, and its probability is called alpha (α). One must specify the level of significance before the experiment in order to be able to interpret it as α. The same holds for beta (β), which is the rate of rejecting the alternative hypothesis H_2 if it is correct (Type-II error). Here we get the second classical interpretation of the level of significance: the error rate α, which is determined before the experiment, albeit not by mere convention, but by cost-benefit calculations that strike a balance between α, β, and sample size n. For instance, if $\alpha = \beta = 0.10$, then it does not matter whether the exact level of significance is 0.06 or 0.001. The level of significance has no influence on α.

3.3. Level of significance = exact level of significance

Fisher had second thoughts about his proposal of a conventional level and stated these most clearly in the 1950s. In his last book, *Statistical Methods and Scientific Inference* (1956, p. 42), Fisher rejected the use of a conventional level of significance and ridiculed this practice, together with the concepts of Type-I and Type-II errors, as "absurdly academic" and originating from "the phantasy of circles rather remote from scientific research" (1956, p. 100). He was referring to mathematicians, specifically to Neyman. In science, Fisher argued, one does not repeat the same experiment again and again, as is assumed in Neyman and Pearson's interpretation of the level of significance as an error rate in the long run. What researchers should do instead, according to Fisher's second thoughts, is publish the *exact level of significance*, say, $p = 0.02$ (not $p < 0.05$). You communicate information; you do not make yes–no decisions.

The basic differences are this: For Fisher, the exact level of significance is a property of the data, that is, a relation between a body of data and a theory. For Neyman and Pearson, α is a property of the test, not of the data. In Fisher's Design, if the result is significant, you reject the null; otherwise you do not draw any conclusion. The decision is asymmetric. In Neyman–Pearson theory, the decision is symmetric. Level of significance and α are not the same thing. For Fisher, these differences were no peanuts. He branded Neyman's position as "childish" and "horrifying [for] the intellectual freedom of the west." Indeed, he likened Neyman to

> Russians [who] are made familiar with the ideal that research in pure science can and should be geared to technological performance, in the comprehensive organized effort of a five-year plan for the nation ... [While] in the U.S. also the great importance of organized technology has I think made it easy to confuse the process appropriate for drawing correct conclusions, with those aimed rather at, let us say, speeding production, or saving money. (Fisher, 1955, p. 70)

It is probably not an accident that Neyman was born in Russia and, at the time of Fisher's comment, had moved to the U.S.

Back to Dr. Publish-Perish and his moral conflict. His superego demands that he specify the level of significance before the experiment. We now understand that his superego's doctrine is part of the Neyman–Pearson theory. His ego personifies Fisher's theory of calculating the exact level of significance from the data, conflated with Fisher's earlier idea of making a yes–no decision based on a conventional level of significance. The conflict

between his superego and his ego is the source of his guilt feelings, but he does not know that. He just has a vague feeling of shame for doing something wrong. Dr. Publish-Perish does not follow any of the three interpretations. Unknowingly, he tries to satisfy all of them, and ends up presenting an exact level of significance as if it were an alpha level, by rounding it up to one of the conventional levels of significance, $p < 0.05$, $p < 0.01$, or $p < 0.001$. The result is not α, nor an exact level of significance. It is the product of an unconscious conflict.

The conflict is institutionalized in the *Publication Manuals of the American Psychological Association*. The fifth edition of the *Manual* (2001, p. 162) has finally added exact levels of significance to an ANOVA (analysis of variance) table, but at the same time retained the $p < 0.05$ and $p < 0.01$ "asterisks" of the null ritual. The manual offers no explanation as to why both are necessary and what they mean (Fidler, 2002). Nor can Dr. Publish-Perish find in it information about the conflicting interpretations of "level of significance" and the origins of his feelings of guilt.

4. Collective illusions

Rituals call for cognitive illusions. Their function is to make the final product, a significant result, appear highly informative, and thereby justify the ritual. Try to answer the following question (Oakes, 1986; Haller and Krauss, 2002):

> Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means *t*-test and your result is significant ($t = 2.7$, d.f. $= 18$, $p = 0.01$). Please mark each of the statements below as "true" or "false." "False" means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.
>
> 1.You have absolutely disproved the null hypothesis (that is, there is no difference between the population means).
>
> [] true/false []
>
> 2. You have found the probability of the null hypothesis being true.
>
> [] true/false []
>
> 3. You have absolutely proved your experimental hypothesis (that there is a difference between the population means).
>
> [] true/false []
>
> 4. You can deduce the probability of the experimental hypothesis being true.
>
> [] true/false []
>
> 5.You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.
>
> [] true/false []

6. You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

[] true/false []

Which statements are in fact true? Recall that a *p*-value is the probability of the observed data (or of more extreme data points), given that the null hypothesis H_0 is true, defined in symbols as $p(D|H_0)$.This definition can be rephrased in a more technical form by introducing the statistical model underlying the analysis (Gigerenzer et al., 1989, chapter 3).

Statements 1 and 3 are easily detected as being false, because a significance test can never disprove the null hypothesis or the (undefined) experimental hypothesis. They are instances of the *illusion of certainty* (Gigerenzer, 2002).

Statements 2 and 4 are also false. The probability $p(D|H_0)$ is not the same as $p(H_0|D)$, and more generally, a significance test does not provide a probability for a hypothesis. The statistical toolbox, of course, contains tools that would allow estimating probabilities of hypotheses, such as Bayesian statistics. Statement 5 also refers to a probability of a hypothesis. This is because if one rejects the null hypothesis, the only possibility of making a wrong decision is if the null hypothesis is true. Thus, it makes essentially the same claim as Statement 2 does, and both are incorrect.

Statement 6 amounts to the replication fallacy (Gigerenzer, 1993, 2000). Here, $p = 1\%$ is taken to imply that such significant data would reappear in 99% of the repetitions. Statement 6 could be made only if one knew that the null hypothesis was true. In formal terms, $p(D|H_0)$ is confused with $1 - p(D)$.

To sum up, all six statements are incorrect. Note that all six err in the same direction of wishful thinking: They make a *p*-value look more informative than it is.

Haller and Krauss (2002) posed the above question to 30 statistics teachers, including professors of psychology, lecturers, and teaching assistants, 39 professors and lecturers of psychology (not teaching statistics), and 44 psychology students. Teachers and students were from the psychology departments at six German universities. Each statistics teacher taught null hypothesis testing, and each student had successfully passed one or more statistics courses in which it was taught. Fig. 1 shows the results.

None of the students noticed that all of the statements were wrong; every student endorsed one or more of the illusions about the meaning of a *p*-value. Perhaps these students lacked the right genes for statistical thinking? Or they did not pay attention to their teachers, and were simply lucky in passing the exams? The results, however, indicate a different explanation. The students inherited the illusions from their teachers. Ninety percent of the professors and lecturers believed one or more of the six statements to be correct. Most surprisingly, 80% of the statistics teachers shared illusions with their students. Note that one does not need to be a brilliant mathematician to answer the question "What does a significant result mean?" One only needs to understand that a *p*-value is the probability of the data (or more extreme data), given that the H_0 is true.The most frequent illusion was Statement 5, endorsed by about 70% of all three groups. In an earlier study with academic psychologists in the UK (Oakes, 1986) as many as 86% thought that this statement was true. The replication fallacy (Statement 6) was the second most frequent illusion, believed to be true by about half of the teachers and 37% of those who taught statistics. The corresponding figure for the UK psychologists was

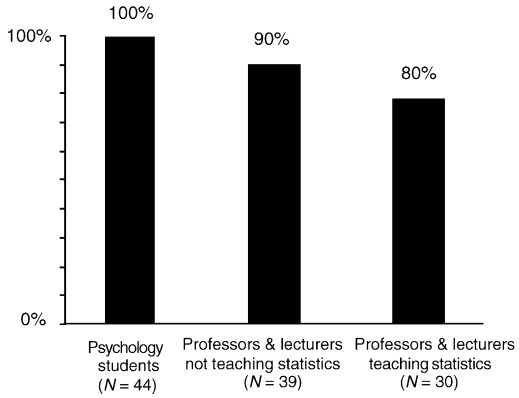

Fig. 1. The amount of delusions about the meaning of "$p = 0.01$." The percentages refer to the participants in each group who endorsed one or more of the six false statements (see Gigerenzer et al., 2004; Haller and Krauss, 2002).

60%. About 60% of the students and one third of each of the teacher groups believed that one can deduce the probability that the experimental hypothesis is true from the p-value (Statement 4). In Oakes' study, two thirds of British academic psychologists believed this. On average, students endorsed 2.5 illusions, their professors and lecturers 2.0 illusions, and those who taught significance testing endorsed 1.9 illusions (Gigerenzer et al., 2004; Haller and Krauss, 2002). All in all, the German professors and lecturers did somewhat better than the British academic psychologists studied earlier by Oakes (1986), yet the number of illusions they held remains breathtaking. Falk and Greenbaum (1995) added the right alternative ("none of the statements is correct") and also made Israeli students read Bakan's (1966) classical article, which warns of these illusions. Nevertheless, 87% of the students opted for one or several illusions. A global fantasy seems to travel by cultural transmission from teacher to student.

If students "inherited" the illusions from their teachers, where did the teachers acquire them? The answer is right there in the first textbooks introducing psychologists to null hypothesis testing more than 50 years ago. Guilford's *Fundamental Statistics in Psychology and Education,* first published in 1942, was probably the most widely read textbook in the 1940s and 1950s. Guilford suggested that hypothesis testing would reveal the probability that the null hypothesis is true. "If the result comes out one way, the hypothesis is probably correct, if it comes out another way, the hypothesis is probably wrong" (p. 156). Guilford's logic wavered back and forth between correct and incorrect statements, and ambiguous ones that can be read like Rorschach inkblots. He used phrases such as "we obtained directly the probabilities that the null hypothesis was plausible" and "the probability of extreme deviations from chance" interchangeably for the level of significance. Guilford is no exception. He marked the beginning of a genre of statistical texts that vacillate between the researchers' desire for probabilities of hypotheses and what significance testing can actually provide. For instance, within three pages of text, Nunally (1975, pp. 194–196; italics in the original) used all of the following statements to explain what a significant result such as 5% actually means:

- "the probability that an observed difference is real"
- "the *improbability* of observed results being due to error"
- "the *statistical confidence* . . . with odds of 95 out of 100 that the observed difference will hold up in investigations"
- "the danger of accepting a statistical result as real when it is actually due only to error"
- the degree to which experimental results are taken "seriously"
- the degree of "faith [that] can be placed in the reality of the finding"
- "the investigator can have 95% confidence that the sample mean actually differs from the population mean"
- "if the probability is low, the null hypothesis is improbable"
- "all of these are different ways to say the same thing"

The poor students who read these explanations! They likely misattribute the author's confusion to their own lack of statistical intelligence. This state of bewilderment will last as long as the ritual continues to exist. Today's students still encounter oracular statements in the most-widely read texts: "Inferential statistics indicate the probability that the particular sample of scores obtained are actually related to whatever you are attempting to measure or whether they could have occurred by chance" (Gerrig and Zimbardo, 2002, p. 44).

Early authors promoting the error that the level of significance specified the probability of hypothesis include Anastasi (1958, p. 11), Ferguson (1959, p. 133), and Lindquist (1940, p. 14). But the belief has persisted over decades: for instance, in Miller and Buckhout (1973; statistical appendix by Brown, p. 523), and in the examples collected by Bakan (1966), Pollard and Richardson (1987), Gigerenzer (1993), Mulaik et al. (1997), and Nickerson (2000). I sometimes hear that if the associated illusions were eliminated, the null ritual would emerge as a meaningful method. As I mentioned before, in contrast, I believe that some degree of illusion is necessary to keep the null ritual alive, and the empirical evidence supports this conjecture (e.g., Lecoutre et al., 2003; Tversky and Kahneman, 1971). Without illusions, the ritual would be easily recognized for what it is.

5. An editor with guts

Everyone seems to have an answer to this question: Who is to blame for the null ritual? Always someone else. A smart graduate student told me that he did not want problems with his thesis advisor. When he finally got his Ph.D. and a post-doc, his concern was to get a real job. Soon he was an assistant professor at a respected university, but he still felt he could not afford statistical thinking because he needed to publish quickly to get tenure. The editors required the ritual, he apologized, but after tenure, everything would be different and he would be a free man. Years later, he found himself tenured, but still in the same environment. And he had been asked to teach a statistics course, featuring the null ritual. He did. As long as the editors of the major journals punish statistical thinking, he concluded, nothing will change.

Blaming editors is not entirely unfounded. For instance, the former editor of the *Journal of Experimental Psychology*, Melton (1962), insisted on the null ritual in his editorial and

also made it clear that he wants to see $p < 0.01$, not just $p < 0.05$. In his editorial, he produced the usual illusions, asserting that the lower the p-value, the higher the confidence that the alternative hypothesis is true, and the higher the probability that a replication will find a significant result. Nothing beyond p-values was mentioned; precise hypotheses, good descriptive statistics, confidence intervals, effect sizes, and power did not appear in the editor's definition of good research. A small p-value was the hallmark of excellent experimentation, a convenient yardstick for whether or not to accept a paper at a time when the number of journals, articles, and psychologists had skyrocketed.

There was resistance. The Skinnerians founded a new journal, the *Journal of the Experimental Analysis of Behavior,* in order to be able to publish their kind of experiments (Skinner, 1984, p. 138). Similarly, one reason for launching the *Journal of Mathematical Psychology* was to escape the editors' pressure to routinely perform null hypothesis testing. One of its founders, R.D. Luce (1988), called this practice a "mindless hypothesis testing in lieu of doing good research: measuring effects, constructing substantive theories of some depth, and developing probability models and statistical procedures suited to these theories" (p. 582).

Should we blame the editors? The story of Geoffrey Loftus, editor of *Memory and Cognition*, however, suggests that the truth is not as simple as that. In 1991, Loftus reviewed *The Empire of Chance* (Gigerenzer et al., 1989), in which we presented one of the first analyses of how psychologists jumbled ideas of Fisher and Neyman–Pearson into one hybrid logic. When Loftus became editor-elect of *Memory and Cognition*, he made it clear in his editorial that he did not want authors to submit papers in which p, t, or F-values had been mindlessly calculated and reported (Loftus, 1993). Rather, his guideline was: "By default, data should be conveyed as a figure depicting sample means *with associated standard errors and/or, where appropriate, standard deviations*"(p. 3; emphasis in the original). His policy encouraged researchers to use proper descriptive statistics, and freed them from the pressure to test null hypotheses and make yes–no decisions whose relevance are obscure. I admire Loftus for the courage to take such a step.

When I met Loftus during his editorship, I asked him how his crusade was going. Loftus bitterly complained about the many researchers who stubbornly refused the opportunity and insisted on their p-values and yes–no decisions. How much success did he have over the years?

Loftus was preceded as editor by Margaret Jean Intons-Petersen, who commenced in 1990. In her incoming editorial, she mentioned the use of descriptive statistics including variability estimates, but emphasized the usual significance tests. During her term, 53% of the articles relied exclusively on the null ritual (Finch et al., 2004). Under Loftus, who served as the editor from 1994 to 1997, this proportion decreased to 32%. During the term of Loftus' successor, Morton Ann Gernsbacher (1998), who did not comment on statistical procedures or on Loftus' recommendations in her editorial, the proportion rose again to about half, reaching a new high of 55% in 2000.

The far majority of the remaining articles also relied on the null ritual but provided some additional information, such as figures with means, standard errors, standard deviations, or confidence intervals. Loftus' recommendation to provide this information without performing the null ritual was followed in only 6% of the articles during his editorship, and in only one (!) case in the years before and after (Finch et al., 2004). Before Loftus, only 8% of the

articles provided figures with error bars and/or reported confidence intervals, and amongst these, in every second case it was left unclear what the bars represented—standard errors, standard deviations, confidence intervals? Loftus brought this proportion up to 45%, and that of unclear error bars down (Finch et al., 2004). But under his successor, the proportion decreased again to 27%, and that of the unclear bars rose.

Loftus reported that many researchers exhibited deep anxiety at the prospect of abandoning their *p*-values, confused standard errors with standard deviations, and had no idea how to compute a confidence interval based on their ANOVA packages. Looking back, he estimated that he requested approximately 300 confidence intervals, and probably computed about 100 himself (Finch et al., 2004). Did Loftus's experiment have the desired impact? During his editorship, he succeeded in reducing reliance on the null ritual; afterwards, the effect declined. Whether his example has a long-term impact is an open question. Loftus was ahead of his time, and I can only hope that his admirable experiment will eventually inspire other editors.

At issue here is the importance of good descriptive and exploratory statistics rather than mechanical hypothesis testing with yes–no answers. Good descriptive statistics (as opposed to figures without error bars, or unclear error bars, and routine aggregate instead of individual analysis, for example) is necessary and mostly sufficient. Note that in scientific problems, the relevance of optimization procedures such as Neyman–Pearson decision theory is notoriously unclear. For instance, unlike in quality control, experimental subjects are rarely randomly sampled from a specified population. Thus, it is unclear for which population the inference from a sample should be made, and "optimal" yes–no decisions are of little relevance. The attempt to give an "optimal" answer to the wrong question has been called "Type-III error." The statistician John Tukey (e.g., 1969) argued for a change in perspective: An appropriate answer to the right problem is better than an optimal answer to the wrong problem (Perlman and Wu, 1999). Neither Fisher's null hypothesis testing nor Neyman–Pearson decision theory can answer most scientific problems. The issue of optimizing versus satisficing is equally relevant for research on bounded rationality and fast and frugal heuristics (Gigerenzer et al., 1999; Todd and Gigerenzer, 2000).

6. The superego, the ego, and the id

Why do intelligent people engage in statistical rituals rather than in statistical thinking? Every person of average intelligence can understand that $p(D|\mathrm{H})$ is not the same as $p(\mathrm{H}|D)$. That this insight fades away when it comes to hypothesis testing suggests that the cause is not intellectual but social and emotional. Here is a hypothesis (Acree, 1978; Gigerenzer, 1993): The conflict between statisticians, both suppressed by and inherent in the textbooks, has become internalized in the minds of researchers. The statistical ritual is a form of conflict resolution, like compulsive hand washing, which makes it resistant to arguments. To illustrate this thesis, I use the Freudian unconscious conflicts as an analogy (Fig. 2).

The Neyman–Pearson theory serves as the superego of Dr. Publish-Perish. It demands in advance that alternative hypotheses, alpha, and power to calculate the sample size necessary be specified precisely, following the frequentist doctrine of repeated random sampling (Neyman, 1957). The superego forbids the interpretation of levels of significance as the

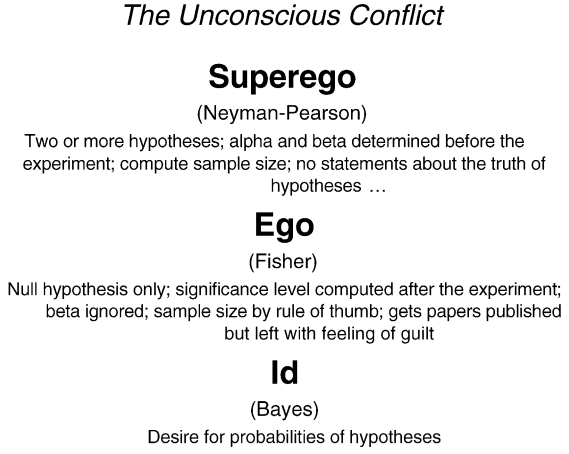

Fig. 2. A Freudian analogy for the unconscious conflict between statistical ideas in the minds of researchers.

degree of confidence that a particular hypothesis is true or false. Hypothesis testing is about what to do, that is, one acts as if a hypothesis were true or false, without necessarily believing that it is true or false.

The Fisherian theory of null hypothesis testing functions as the ego. The ego gets things done in the laboratory and papers published. Levels of significance are computed after the experiment, the power of the test is ignored, and the sample size is determined by a rule of thumb. The ego does not state its research hypothesis in a precise way but at best in form of a directional prediction, yet does not hesitate to claim support for it by rejecting a null hypothesis. The ego makes abundant epistemic statements about its confidence in particular hypotheses. But it is left with feelings of guilt and shame for having violated the rules.

The Bayesian view forms the id. Its goal is a statement about the probabilities of hypotheses, which is censored by both the purist superego and the pragmatic ego. However, these probabilities are exactly what the id wants, after all. It gets its way by blocking the intellect from understanding that $p(D|\mathrm{H})$ is not the same as $p(\mathrm{H}|D)$. This enables wishful thinking.

The Freudian analogy brings the anxiety and the feelings of guilt into the foreground. It seems as if the raging personal and intellectual conflicts between Fisher and Neyman–Pearson, and between these frequentists and the Bayesians were projected into an "intrapsychic" conflict in the minds of researchers. In Freudian theory, ritual is a way of resolving unconscious conflict, but at considerable costs.

7. Meehl's conjecture

Paul Meehl, a brilliant clinical psychologist with a broad interest in the philosophy of science, was one of those who blamed Fisher for the decline of statistical thinking in psychology. "Sir Ronald has befuddled us, mesmerized us, and led us down the primrose path. I believe the almost universal reliance on merely refuting the null hypothesis … is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst

things that ever happened in the history of psychology" (Meehl, 1978, p. 817). Meehl is a bit harsh on blaming Fisher rather the null ritual; recall that Fisher also proposed other statistical tools, and in the 1950s, he thought of null hypothesis testing as adequate only for situations in which we know nothing or little. Meehl (1978) made a challenging prediction concerning null hypothesis tests in nonexperimental settings, where random assignment to treatment and control group is not possible, due to ethical or practical constraints. It can be summarized as follows:

Meehl's conjecture:

In nonexperimental settings with large sample sizes, the probability of rejecting the null hypothesis of nil group differences in favor of a directional alternative is about 0.50.

Isn't that good news? We guess that X is larger than Y—and we get it right half of the time. For instance, if we make up the story that Protestants have a higher memory span than Catholics, slower reaction times, smaller shoe size, and higher testosterone levels, each of these hypotheses has about a 50% chance of being accepted by a null hypothesis test. If we do not commit to the direction and just guess that X and Y are different, we get it right virtually 100% of the time. Meehl reasoned that in the real world—as opposed to experimental settings—the null hypothesis ("nil" as defined by the null ritual, not by Fisher) is always wrong. Some difference exists between any natural groups. Therefore, with sufficient statistical power, one will almost always find a significant result. If one randomly guesses the direction of the difference, it follows that one will be correct in about 50% of the cases (with a unidirectional alternative hypothesis, one will be correct in about 100% of them).

Niels Waller (2004) set out to test Meehl's conjecture empirically. He had access to the data of more than 81,000 individuals who had completed the 567 items of the Minnesota Multiphase Personality Inventory—Revised (MMPI-2). The MMPI-2 asks people about a broad range of contents, including health, personal habits, attitudes toward sex, and extreme manifestations of psychopathology. Imagine a gender theorist who has concocted a new theory that predicts directional gender differences, that is, women will score higher on some item than men, or vice versa. Can we predict the probability of rejecting the null hypothesis in favor of the new theory? According to Meehl's conjecture, it is about 50%. In Waller's simulation, the computer picked the first of the 511 items of the MMPI-2 (excluding 56 for their known ability to discriminate between the sexes), determined randomly the direction of the alternative hypothesis, and computed whether the difference was significant in the predicted direction. This procedure was repeated with all 511 items. The result: 46% of the predictions were confirmed, often with very impressive p-values. Many of the item mean differences were 50–100 times larger than the associated standard errors!

These empirical results support Meehl's conjecture, consistent with earlier findings by Bakan (1966) and Meehl himself. A bit of statistical thinking can make the logic of the conjecture transparent to an undergraduate. Yet one can find experienced researchers who proudly report that they have studied several hundreds or even thousands of subjects and found a highly significant mean difference in the predicted direction, say $p < 0.0001$. How big this effect is, however, is not reported in some of these articles. The combination of large sample size and low p-values is of little value in itself.

The general problem addressed by Meehl is the inattention to effect sizes in the null ritual. Effect sizes have been discussed by Cohen (1988) and Rosenthal and Rubin (1982). The Task Force on Statistical Inference (TFSI) of the *American Psychological Association* (Wilkinson and TFSI, 1999) recommended reporting effect sizes (theoretical ones as in Neyman–Pearson theory, or empirical ones) as essential. The fifth edition of the *Publication Manual of the American Psychological Association* (2001) followed up this recommendation, although only half-heartedly. In the examples given, effect sizes are either not included or not explained and interpreted (Fidler, 2002).

Without a theoretical effect size, the statistical power of a test cannot be computed. In 1962, Jacob Cohen reported that the experiments published in a major psychology journal had, on average, only a fifty-fifty chance of detecting a medium-sized effect if there was one. That is, the statistical power was as low as 50%. This result was widely cited, but did it change researchers' practice? Sedlmeier and Gigerenzer (1989) checked the studies in the same journal, 24 years later, a time period that should allow for change. Yet only 2 out of 64 researchers mentioned power, and it was never estimated. Unnoticed, the average power had actually decreased (researchers now used alpha adjustment, which shrinks power). Thus, if there had been an effect of a medium size, the researchers would have had a better chance of finding it by throwing a coin rather than conducting their time-consuming, elaborate, and expensive experiments. In the years 2000–2002, amongst some 220 empirical articles, there were finally 9 researchers who computed the power of their tests (Gigerenzer et al., 2004). Forty years after Cohen, there is a first sign of change. The fourth edition of the *Publication Manual of the American Psychological Association* (1994) was the first to recommend that researchers take power seriously, and the fifth edition (2001) repeated this advice. Yet despite an abundance of examples for how to report *p*-values, the manual still does not include any examples of reporting power (Fidler, 2002).

8. Feynman's conjecture

The routine reliance on the null ritual discourages not only statistical thinking but also theoretical thinking. One does not need to specify one's hypothesis, nor any challenging alternative hypothesis. There is no premium on "bold" hypotheses, in the sense of Karl Popper or Bayesian model comparison (MacKay, 1995). In many experimental papers in social and cognitive psychology, there is no theory in shooting distance, but only surrogates such as redescription of the results (Gigerenzer, 2000, chapter 14). The sole requirement is to reject a null that is identified with "chance." Statistical theories such as Neyman–Pearson theory and Wald's theory, in contrast, begin with two or more statistical hypotheses.

In the absence of theory, the temptation is to look first at the data and then see what is significant. The physicist Richard Feynman (1998, pp. 80–81) has taken notice of this misuse of hypothesis testing. I summarize his argument.

Feynman's conjecture:

To report a significant result and reject the null in favor of an alternative hypothesis is meaningless unless the alternative hypothesis has been stated before the data was obtained.

When he was a graduate student at Princeton, Feynman got into an argument with a researcher in the psychology department. The researcher had designed an experiment, in which rats ran in a T-maze. The rats did not behave as predicted. Yet the researcher noticed something else, that the rats seem to alternate, first right, then left, then right again, and so on. He asked Feynman to calculate the probability under the null hypothesis (chance) that this pattern would be obtained. On this occasion, Feynman (1998) learned about the 5% level:

> And it's a general principle of psychologists that in these tests they arrange so that the odds that the things that happen happen by chance is small, in fact, less than one in twenty. . . . And then he ran to me, and he said, "Calculate the probability for me that they should alternate, so that I can see if it is less than one in twenty." I said, "It probably is less than one in twenty, but it doesn't count." He said, "Why?" I said, "Because it doesn't make any sense to calculate after the event. You see, you found the peculiarity, and so you selected the peculiar case." . . . If he wants to test this hypothesis, one in twenty, he cannot do it from the same data that gave him the clue. He must do another experiment all over again and then see if they alternate. He did, and it didn't work." (pp. 80–81)

Feynman's conjecture is again and again violated by routine significance testing, where one looks at the data to see what is significant. Statistical packages allow every difference, interaction, or correlation against chance to be tested. They automatically deliver ratings of "significance" in terms of stars, double stars, and triple stars, encouraging the bad after-the-fact habit. The general problem Feynman addressed is known as overfitting. Fitting a model to data that is already obtained is not sound hypothesis testing, even if the resulting explained variance, or R^2, is impressive. The reason is that one does not know how much noise one has fitted, and the more adjustable parameters one has, the more noise one can fit. Psychologists habitually fit rather than predict, and rarely test a model on new data, such as by cross-validation (Roberts and Pashler, 2000). Fitting per se has the same problems as story telling after the fact, which leads to a "hindsight bias" (Hoffrage et al., 2000). The true test of a model is to fix its parameters on one sample, and to test it in a new sample. Then it turns out that predictions based on simple heuristics can be more accurate than routine multiple regressions (Czerlinski et al., 1999). Less can be more. The routine use of linear multiple regression exemplifies another mindless use of statistics.

9. The dawn of statistical thinking

Rituals seem to be indispensable for the self-definition of social groups and for transitions in life, and there is nothing wrong with them. However, they should be the subject rather than the procedure of social sciences. Elements of social rituals include (i) the repetition of the same action, (ii) a focus on special numbers or colors, (iii) fears about serious sanctions for rule violations, and (iv) wishful thinking and delusions that virtually eliminate critical thinking (Dulaney and Fiske, 1994). The null ritual has each of these four characteristics: the same procedure is repeated again and again; the magical 5% number; fear of sanctions by editors or advisors, and wishful thinking about the outcome, the *p*-value, which blocks researchers' intelligence.

We know but often forget that the problem of inductive inference has no single solution. There is no uniformly most powerful test, that is, no method that is best for every problem. Statistical theory has provided us with a toolbox with effective instruments, which require judgment about when it is right to use them. When textbooks and curricula begin to teach the toolbox, students will automatically learn to make judgments. And they will realize that in many applications, a skilful and transparent descriptive data analysis is sufficient, and preferable to the application of statistical routines chosen for their complexity and opacity. Judgment is part of the art of statistics.

To stop the ritual, we also need more guts and nerves. We need some pounds of courage to cease playing along in this embarrassing game. This may cause friction with editors and colleagues, but it will in the end help them to enter the dawn of statistical thinking.

References

Acree, M.C., 1978. Theories of Statistical Inference in Psychological Research: A Historicocritical Study. Dissertation. University Microfilms International H790 H7000, Ann Arbor, MI.

American Psychological Association, 1974. Publication Manual, 2nd ed., 3rd ed., 1983; 4th ed., 1994; 5th ed., 2001. Garamond/Pridemark Press, Baltimore, MD.

Anastasi, A., 1958. Differential psychology, 3rd ed. Macmillan, New York.

Anderson, D.R., Burnham, K.P., Thompson, W.L., 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64, 912–923.

Bakan, D., 1966. The test of significance in psychological research. Psychological Bulletin 66, 423–437.

Chow, S.L., 1998. Précis of "Statistical significance: rationale, validity, and utility". Behavioral and Brain Sciences 21, 169–239.

Cohen, J., 1962. The statistical power of abnormal-social psychological research: a review. Journal of Abnormal and Social Psychology 65, 145–153.

Cohen, J., 1988. Statistical power analysis for the behavioral sciences, 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ.

Czerlinski, J., Gigerenzer, G., Goldstein, D.G., 1999. How good are simple heuristics? In: Gigerenzer, G., Todd, P.M., the ABC Reading Group, Simple Heuristics That Make Us Smart. Oxford University Press, New York, pp. 97–118.

Dulaney, S., Fiske, A.P., 1994. Cultural rituals and obsessive-compulsive disorder: is there a common psychological mechanism? Ethos 22, 243–283.

Falk, R., Greenbaum, C.W., 1995. Significance tests die hard. Theory and Psychology 5, 75–98.

Ferguson, L., 1959. Statistical Analysis in Psychology and Education. McGraw-Hill, New York.

Feynman, R., 1998. The Meaning of it All: Thoughts of a Citizen-Scientist. Perseus Books, Reading, MA, pp. 80–81.

Fidler, F., 2002. The fifth edition of the APA Publication Manual: why its statistics recommendations are so controversial. Educational and Psychological Measurement 62, 749–770.

Finch Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., et al. 2004. Reform of statistical inference in psychology: the case of memory and cognition. Behavior Research Methods, Instruments and Computers 36, 312–324.

Fisher, R.A., 1935. The design of experiments, 5th ed., 1951; 7th ed., 1960; 8th ed., 1966. Oliver & Boyd, Edinburgh.

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

Fisher, R.A., 1956. Statistical Methods and Scientific Inference. Oliver & Boyd, Edinburgh.

Gernsbacher, M.A., 1998. Editorial comment. Memory and Cognition 26, 1.

Gerrig, R.J., Zimbardo, P.G., 2002. Psychology and Life, 16th ed. Allyn and Bacon, Boston.

Gigerenzer, G., 1987. Probabilistic thinking and the fight against subjectivity. In: Krüger, L., Gigerenzer, G., Morgan, M. (Eds.), The Probabilistic Revolution. Vol. II: ideas in the Sciences. MIT Press, Cambridge, MA, pp. 11–33.

Gigerenzer, G., 1993. The superego, the ego, and the id in statistical reasoning. In: Keren, G., Lewis, C. (Eds.), A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues. Erlbaum, Hillsdale, NJ, pp. 311–339.

Gigerenzer, G., 2000. Adaptive Thinking: Rationality in the Real World. Oxford University Press, New York.

Gigerenzer, G., 2002. Calculated Risks: How to Know When Numbers Deceive You. Simon & Schuster, New York (UK edition: Reckoning with Risk: Learning to Live with Uncertainty. Penguin, London).

Gigerenzer, G., Krauss, S., Vitouch, O., 2004. The null ritual: What you always wanted to know about null hypothesis testing but were afraid to ask. In: Kaplan, D. (Ed.), Handbook on Quantitative Methods in the Social Sciences. Sage, Thousand Oaks, CA, pp. 389–406.

Gigerenzer, G., Murray, D.J., 1987. Cognition as Intuitive Statistics. Erlbaum, Hillsdale, NJ.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Krüger, L., 1989. The Empire of Chance. How Probability Changed Science and Every Day Life. Cambridge University Press, Cambridge, UK.

Gigerenzer, G., Todd, P.M., The ABC Research Group, 1999. Simple Heuristics that Make Us Smart. Oxford University Press, New York.

Guilford, J.P., 1942. Fundamental Statistics in Psychology and Education, 3rd ed., 1956; 6th ed., 1978 (with Fruchter, B). McGraw-Hill, New York.

Haller, H., Krauss, S., 2002. Misinterpretations of significance: a problem students share with their teachers? Methods of Psychological Research—Online [On-line serial], 7, 1–20. Retrieved June 10, 2003, from http://www.mpr-online.de.

Hoffrage, U., Hertwig, R., Gigerenzer, G., 2000. Hindsight bias: a by-product of knowledge updating? Journal of Experimental Psychology: Learning, Memory, and Cognition 26, 566–581.

Intons-Peterson, M.J., 1990. Editorial. Memory and Cognition 18, 1–2.

Lecoutre, M.P., Poitevineau, J., Lecoutre, B., 2003. Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. International Journal of Psychology 38, 37–45.

Lindquist, E.F., 1940. Statistical Analysis in Educational Research. Houghton Mifflin, Boston.

Loftus, G.R., 1991. On the tyranny of hypothesis testing in the social sciences. Contemporary Psychology 36, 102–105.

Loftus, G.R., 1993. Editorial comment. Memory and Cognition 21, 1–3.

Luce, R.D., 1988. The tools-to-theory hypothesis. Review of G. Gigerenzer and D.J. Murray, "Cognition as intuitive statistics". Contemporary Psychology 33, 582–583.

MacKay, D.J., 1995. Probable networks and plausible predictions: a review of practical Bayesian methods for supervised neural networks. Network: Computation in Neural Systems 6, 469–505.

McCloskey, D.N., Ziliak, S., 1996. The standard error of regression. Journal of Economic Literature 34, 97–114.

Meehl, P.E., 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology 46, 806–834.

Melton, A.W., 1962. Editorial. Journal of Experimental Psychology 64, 553–557.

Miller, G.A., Buckhout, R., 1973. Psychology: The Science of Mental Life. Harper & Row, New York.

Morrison, D.E., Henkel, R.E., 1970. The Significance Test Controversy. Aldine, Chicago.

Mulaik, S.A., Raju, N.S., Harshman, R.A., 1997. There is a time and a place for significance testing. In: Harlow, L.L., Mulaik, S.A., Steiger, J.H. (Eds.), What if there were no significance tests? Erlbaum, Mahwah, NJ, pp. 65–115.

Neyman, J., 1950. First Course in Probability and Statistics. Holt, New York.

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

Nickerson, R.S., 2000. Null hypothesis significance testing: a review of an old and continuing controversy. Psychological Methods 5, 241–301.

Nunally, J.C., 1975. Introduction to Statistics for Psychology and Education. McGraw-Hill, New York.

Oakes, M., 1986. Statistical Inference: A Commentary for the Social and Behavioral Sciences. Wiley, NY.

Perlman, M.D., Wu, L., 1999. The emperor's new tests. Statistical Science 14, 355–381.

Pollard, P., Richardson, J.T.E., 1987. On the probability of making Type I errors. Psychological Bulletin 102, 159–163.

Roberts, S., Pashler, H., 2000. How persuasive is a good fit? A comment on theory testing. Psychological Review 107, 358–367.

Rosenthal, R., Rubin, D.R., 1982. Comparing effect sizes of independent studies. Psychological Bulletin 92, 500–504.

Rucci, A.J., Tweney, R.D., 1980. Analysis of variance and the "second discipline" of scientific psychology: a historical account. Psychological Bulletin 87, 166–184.

Sedlmeier, P., Gigerenzer, G., 1989. Do studies of statistical power have an effect on the power of studies? Psychological Bulletin 105, 309–316.

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

Skinner, B.F., 1972. Cumulative record. Appleton-Century-Crofts, New York.

Skinner, B.F., 1984. A Matter of Consequences. New York University Press, New York.

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

Todd, P.M., Gigerenzer, G., 2000. Précis of simple heuristics that make us smart. Behavioral and Brain Sciences 23, 727–780.

Tukey, J.W., 1969. Analyzing data: sanctification or detective work? American Psychologist 24, 83–91.

Tversky, A., Kahneman, D., 1971. Belief in the law of small numbers. Psychological Bulletin 76, 105–110.

Waller, N.G., 2004. The fallacy of the null hypothesis in soft psychology. Applied and Preventive Psychology 11, 83–86.

Wilkinson, L., The Task Force on Statistical Inference, 1999. Statistical methods in psychology journals: guidelines and explanations. American Psychologist, 54, 594–604.