International Statistical Review (2001), **69**, 3, 399–417, Printed in The Netherlands
© International Statistical Institute

Uses, Abuses and Misuses of Significance Tests in the Scientific Community: Won't the Bayesian Choice be Unavoidable?

Bruno Lecoutre[1,2], Marie-Paule Lecoutre[2] and Jacques Poitevineau[3,2]

[1]*Laboratoire de Mathématiques Raphaël Salem, UMR 6085 C.N.R.S. et Université de Rouen, Mathématiques, Site Colbert, 76821 Mont-Saint-Aignan Cedex, France*
E-mail: bruno.lecoutre@univ-rouen.fr.
[2]*ERIS, Laboratoire Psy. Co, EA 1780, Université de Rouen, UFR Psychologie, Sociologie, Sciences de l'Éducation, 76821 Mont-Saint-Aignan Cedex, France*
E-mail: marie-paule.lecoutre@univ-rouen.fr
[3]*LCPE, INaLF, FRE2173 C.N.R.S., 44 rue de l'Amiral Mouchez, 75014 Paris, France*
E-mail: jacques.poitevineau@ivry.cnrs.fr

Summary

The current context of the "significance test controversy" is first briefly discussed. Then experimental studies about the use of null hypothesis significance tests by scientific researchers and applied statisticians are presented. The misuses of these tests are reconsidered as judgmental adjustments revealing researchers' requirements towards statistical inference. Lastly alternative methods are considered. Consequently we automatically ask ourselves "won't the Bayesian choice be unavoidable?"

Key words: Bayesian Methods; Confidence intervals; Experimental data analysis; Fisher; Significance test controversy; Statistical inference.

> *"Habit is habit and not to be flung out of the window by any man, but coaxed downstairs a step at a time."* (Mark Twain)

1 Introduction

Experimental research is facing a paradoxical situation. On the one hand, Null Hypothesis Significance Testing (NHST) is required in most scientific publications as an unavoidable norm. NHST is used to strengthen data and convince the community of the value of the results. Furthermore it often appears as a label of scientificness. But on the other hand, NHST leads to innumerable misinterpretations and misuses. Moreover, from the outset (Boring, 1919; Tyler, 1931; Berkson, 1938; etc.), NHST has been subject to intense criticism. Its use has been explicitly denounced by the most eminent and most experienced scientists, both on theoretical and methodological grounds, not to mention the sharp controversy that opposed Fisher to Neyman and Pearson on the very foundations of statistical inference. In the sixties there was more and more criticism, especially in the behavioral and social sciences, denouncing the shortcomings of NHST and demonstrating its inadequacy in

experimental data analysis. Most of these criticisms have been published in experimental journals so they could hardly remain unknown to most researchers. Findings from statistical re-analysis of published results and from experimental investigations of judgments carried out by researchers in situations of statistical inference reveal that the practice of NHST entails considerable distortions. This is especially true in the designing and monitoring of experiments, and in the selection and presentation of results.

So the time has come to reach a consensus on procedures that bypass the common misuses of NHST, while at the same time respecting its role of "an aid to judgment" which "should not be confused with automatic acceptance tests, or 'decision functions'." (Fisher, 1990/1925, page 128). This agreement should meet scientists' demands, in particular the need for objective statements and the need for procedures on effect sizes. Undoubtedly, there is increasing acceptance that Bayesian inference can be ideally suited for this purpose. Moreover, the Bayesian school is progressively becoming the dominant school in mathematical statistics (see e.g., Berger, 1985; Bernardo & Smith, 1994; Robert, 1994; Schervish, 1995) and sooner or later it will have determining implications in teaching both students and scientific researchers. A not unreasonable belief is to anticipate the evolution of statistics and to think that in the future Bayesian inference will be the dominant approach. Lindley (*in* Smith, 1995) even stated: "we [statisticians] will all be Bayesians in 2020, and then we can be a united profession" (page 317).

The present article is divided into three parts. In the first normative part we briefly discuss the current context of the "significance test controversy". In the second descriptive part we investigate from experimental findings, the attitudes of scientific researchers towards significance tests. We examine how misuses of these tests are linked to some major criticisms and can be seen as judgmental adjustments revealing the true needs of researchers towards statistical inference. In the third prescriptive part we examine the impact of alternative solutions. Consequently we automatically ask ourselves: "won't the Bayesian choice be unavoidable?"

2 The Significance Test Controversy

2.1 The Shortcomings of Null Hypothesis Significance Tests

Experimental research can be compared to a game or a fight (Freeman, 1993, used the adjective "gladiatorial"), within which only the significant results win, while nonsignificant ones are (theoretically) only statements of ignorance, and thus perceived as failures. These practices can be seen with Salsburg (1985) as the "religion of statistics" with rites such as the use of the "profoundly mysterious symbols of the religion *NS*, *, **, and *mirabile dictu* ***". On the same lines, Guttman (1983), denounced the "star worshippers" and openly attacked the fact that some scientific journals, and *Science* in particular, consider the significance test as a criterion of scientificness. As a matter of fact, a very frequent error consists in mistaking statistical significance for scientific significance: the more significant a result is, the more scientifically interesting it is, and/or the larger the true effect is. This has been one of the most often denounced errors (Selvin, 1957; Kish, 1959; Bolles, 1962; Reuchlin, 1962; Bakan, 1966; O'Brien & Shapiro, 1968; Gold, 1969; Morrison & Henkel, 1969; Winch & Campbell, 1969, etc.). From a survey of research articles published in three different psychology journals, Craig, Eison & Metze (1976) concluded that "researchers and journal editors as a whole tend to (over)rely on 'significant differences' as the definition of meaningful research" (page 282). This leads to publication biases denounced by many authors (e.g., Tullock, 1959; McNemar, 1960; Bakan 1966). Sterling (1959) actually found as early as 1955 and 1956 that the vast majority of published articles in four randomly selected psychology journals satisfied a minimum criterion of significance. Out of 81 per cent of articles using tests, more than 97 per cent rejected H_0 when considering the major hypotheses. Thirty years later the situation had not evolved (Sterling, Rosenbaum & Weinkam, 1995).

However even in this context of the *significant test dictatorship*, and in spite of continuous warnings, nonsignificant results are also frequently used improperly as *proof of the null hypothesis* in experimental publications (see e.g., Harcum, 1990). So when considering all the hypotheses tested and not only the major ones, about half of the articles published in the 1994 issue of the *Journal of Abnormal Psychology* contained conclusions such as "there is no difference between groups" or "there is no interaction effect" based on nonsignificant tests (Poitevineau, 1997). Furthermore even if "no effect" was understood as "small or negligible effect", standard Bayesian re-analysis clearly demonstrated that such conclusions were in most cases unjustified.

2.2 *The Current State of Controversy*

Until now the social phenomenon of NHST seems to have resisted all warnings. The history of this resistance can be summarized by some especially revealing titles of articles: criticisms—The fallacy of the null hypothesis significance test (Rozeboom, 1960); sentences—The case against statistical significance testing (Carver, 1978); death predictions—The end of the p-value? (Evans, Mills & Dawson, 1988); death that does not occur—Significance tests die hard (Falk & Greenbaum, 1995). If some people like Hogben (1957) recommended definitively abandoning all statistical inference methods, they are a minority. It is clear that statistical inference methods are unavoidable. A handrail is at least necessary to prevent researchers from "getting carried away" by hasty impressionistic generalizations. Many of the authors who have criticized significance testing have therefore proposed solutions. Let us mention in particular, by limiting ourselves to some early references: replications of experiments (Tullock, 1959); analysis of effect sizes (Nunnally, 1960); confidence intervals (Natrella, 1960; Grant, 1962); Bayesian methods and/or likelihood methods (Rozeboom, 1960; Edwards, Lindman & Savage, 1963); power studies (Binder, 1963; Cohen, 1962, 1977); appropriate sample size estimation (Freiman *et al.*, 1978); tests of shifted null hypotheses (Fowler, 1985; Victor, 1987); meta-analyses (Glass, McGaw & Smith, 1981).

Nowadays, academic debates are repetitive and give a discouraging feeling of *déjà-vu*. Moreover many recent papers are replete with ill-informed secondary sources, or ill-considered claims, and first and foremost concerning Fisherian and Bayesian inferences. Unfortunately this confusing controversy, rather than stimulating the interest of experimental scientists, is without doubt detrimental to the impact of new proposals, if not to the image of statistical inference. It is perhaps a major reason why alternative solutions are rarely used in practice and encounter an *inertia* amongst users as well as amongst statisticians, who uphold the use of NHST by relying upon traditions and practices. It is much easier for a scientist to fall back on an automated, socially approved procedure than to look for alternative methods of analysis and risk having his or her paper rejected for publication. In addition there are ways of appeasing one's conscience: "There were far too many studies to plan and too much data to analyze to worry seriously about what the p-values and confidence coefficients produced by the package actually meant." (Breslow, 1990, page 269).

3 The Uses of Null Hypothesis Significance Testing Revisited from Experimental Findings

Experimental research about the use of NHST in the scientific community, and more specifically in psychology, can be divided into two categories.

(1) Firstly by means of multiple choice questions surveys investigated how well psychologists use statistical significance tests (e.g., Zuckerman *et al.*, 1993) or how they perceive them (e.g. Mittag & Thompson, 2000). Most of these surveys induced stereotypical answers and reflected users' theoretical knowledge in statistics more than their own opinions and practices. Therefore they will not be considered any further in this article.

(2) In contrast, research in the second category aimed at studying the *spontaneous* interpretations

of NHST. In one of the first experiments on the use of significance tests (Rosenthal & Gaito, 1963, 1964; Beauchamp & May, 1964; Nelson, Rosenthal & Rosnow, 1986), researchers in psychology were asked to state their degree of belief in the hypothesis of an effect as a function of the associated p-values and sample sizes. The degree of belief decreased when the p-value increased, and was on average approximately an exponential function. However the authors emphasized a *cliff effect* for the .05 level, i.e. "an abrupt drop" in confidence just beyond this level. Although this cliff effect was at most of relatively moderate magnitude, they came to the conclusion that "research decisions to believe or not to believe the null hypothesis (accept not accept) are made in a binary manner based simply on whether p does or does not reach the .05 level". Other experiments where subjects were faced with different possible interpretations of p-values, pointed out the lack of naturalness of the "correct" frequentist interpretation. Indeed, $1 - p$ was most often interpreted as the probability that the alternative hypothesis was true or as evidence of the replicability of the result (Oakes, 1986; Wulff *et al.*, 1987; Scheutz, Andersen & Wulff, 1988; Freeman, 1993; Falk & Greenbaum, 1995). Ironically these "naive" subjects are in good company, with Student ("the probability is 0.9985 [$1 - p$] that [soporific] 2 is the better soporific", Student, 1908, page 21), and again more surprisingly with Neyman himself ("in these conditions [a p-value of 1/15], the odds of 14 to 1 that this loss was caused by seeding [of clouds] do not appear negligible to us", Neyman *et al.*, 1969).

Tversky & Kahneman (1971) initiated more ecological experiments in the general context of research on uncertainty situations. According to these authors, people in these situations develop various heuristics that could explain some misconceptions of NHST by *biases* in probabilistic judgments. For instance, they invoked the *representativeness hypothesis*, according to which the overestimation of the replicability of an experimental result is due to an unjustifiably high degree of confidence that any two samples from the same population resemble each other (Kahneman, Slovic & Tversky, 1982). Oakes (1986) advocated the notion of *significance hypothesis*, according to which a significance test is interpreted in terms of a dichotomy: an effect either "exists" when it is significant, or "does not exist" when it is nonsignificant. He referred to the .05 cliff effect mentioned above as evidence of this hypothesis (page 83). Other relevant references about this approach are Kahneman & Tversky, 1972; Nisbett & Ross, 1981; Gigerenzer & Murray, 1987; Gigerenzer, 1993.

Within the perspective of describing and analyzing the practices and attitudes towards NHST of experienced users, such as scientific researchers or professional applied statisticians, we developed an experimental research project. It aimed at collecting spontaneous answers based on each user's personal experience, which reflect above all his or her own convictions. A guiding principle was to confront subjects with *conflicting* situations. In these situations there was an apparent contradiction or conflict between a given NSHT outcome and other information. For instance a t test for comparing two means was nonsignificant but the descriptive results showed a large observed difference, or the results of one experiment diverge from those of another said to replicate it, etc. (see Lecoutre, 2000). We also studied statistical prediction situations, which for instance consisted in asking subjects to estimate the probability, given a significant result, that this result would be significant once again in a replication of the experiment (Lecoutre & Rouanet, 1993). These situations lead one to examine indirectly how a statistical conclusion is understood, and the confidence that users have in statistical conclusions based on a null hypothesis significance test.

Three experiments carried out in this context are summarized hereafter. They investigate the role of the various "ingredients" that are commonly available in publications for interpreting statistical results. Experiment 1 consisted in a replication of the aforementioned experiment by Rosenthal & Gaito (1963) on the interpretation of p-values. Indeed it is not uncommon to find published papers, especially in some medical journals, that report *nothing but* p-values. Our aim was to identify distinct categories of subjects, possibly corresponding to different conceptions of statistical inference, referring in particular to Neyman–Pearson, Fisher and Bayes. Experiments 2 and 3 were designed to confront subjects with conflicting situations. In all our experiments, subjects were asked to respond

in a spontaneous fashion, without making calculations. It was stressed that the task was in no way a test of knowledge of statistics. Earlier studies which led to very similar findings can be found in Lecoutre (1982, 1983).

The subjects were psychology researchers from various laboratories in France, all with practical experience of processing experimental data. Furthermore Experiment 3 aimed at comparing psychologists and professional statisticians, therefore "expert" subjects in statistics from various pharmaceutical companies. Subjects carried out the task individually. Experiment 1 lasted from 5 to 10 minutes. In the two other experiments, the responses and their justifications were gathered by means of semi-directive interviews. These interviews ranged in length from 15 minutes to half an hour.

3.1 Experiment 1: a Replication of Rosenthal and Gaito

12 p-values (.001, .01, .03, .05, .07, .10, .15, .20, .30, .50, .70, .90) combined with two sample sizes ($n = 10$ and $n = 100$ as in the original experiment) were presented at random, on separate pages of a notebook. It was specified that the test was a Student's t for paired groups. The subjects were asked to state their degree of belief in the hypothesis that "experimental treatment really had an effect". They were asked to tick off a point on a non-graduated segment line of 10 centimeters, from null confidence (left extremity) to full confidence (right extremity of the scale). The subjects' responses were measured in the [0, 1] interval. 18 psychology researchers carried out this experiment.

Results

Although our experiment was conducted about 35 years after the original one and in another country, the average curves appeared to be similar. As in Rosenthal & Gaito's (1963) study, the degree of belief was always greater for $n = 100$ than for $n = 10$. A .05 "cliff effect" was also apparent for the two sample sizes, however the average curves were fairly well fitted by an exponential function. However in actual fact, the study of individual curves revealed that subjects could actually be classified into three clearly distinct categories, the classification being identical for the two curves ($n = 10$ and $n = 100$) of each subject (see Figure 1).

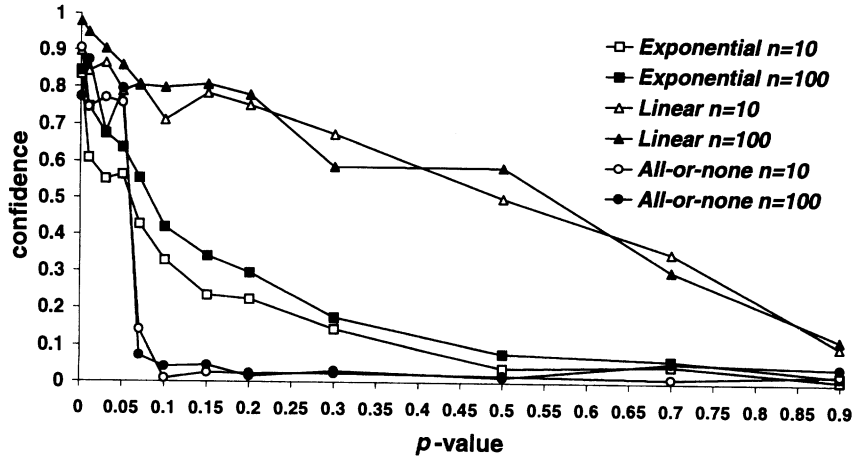

Figure 1. *Experiment 1: confidence in the hypothesis that the experimental treatment really had an effect as a function of the p-value and the sample size n, for each of the three identified groups.*

(1) 10 out of 18 subjects presented a decreasing exponential curve. It was similar to the type of curve often obtained in psychophysics experiments (where psychological properties of objects are related to physical measures), as if these subjects considered the p-values as a physical measure of weight of evidence. (2) 4 subjects presented a negative linear curve. These results were compatible with the common misinterpretation of a p-value as the complement of the probability that the alternate hypothesis is true, which Carver (1978) called the *Valid Research Hypothesis Fantasy*. (3) 4 subjects presented an all-or-none curve with a very high degree of belief when $p \leq 0.05$ and with nearly a null degree of belief otherwise. Only these stepwise curves clearly referred to a decision making attitude.

The larger sample size gave more confidence to the subjects in the first category, whereas all the other subjects had almost the same degree of belief for a given p, whatever the sample size.

To sum up, a major finding of this study was that the attitude of psychology researchers towards p-values was far from being as homogeneous as might be expected. Moreover most of them rated graduated judgments, either exponential or linear, and it was mainly because of a minority of all-or-none respondents (4 out of 18) that an average .05 cliff effect stood. Thus the previous claims about the existence of "an abrupt drop" in a p level just beyond the fateful .05 level (Rosenthal & Gaito, 1963; Nelson, Rosenthal & Rosnow, 1986; Oakes, 1986) should be seriously moderated.

3.2 *Experiment 2: the "Choice-of-Criteria Questionnaire"*

The "choice-of-criteria questionnaire" covered eight cases defined by combining the three types of information most often available when carrying out a statistical analysis of experimental data, namely, the *a priori* expectations, the descriptive procedure results, and the significance test outcome. Two modalities were chosen for each type of information: (i) one expected to find a difference between the two experimental conditions vs. one had no expectations at all; (ii) the difference between the means observed under the two experimental conditions was small vs. large; (iii) the Student's t test was very significant ($p \leq .01$) vs. nonsignificant ($p > .10$). 23 psychology researchers were given this questionnaire. They were asked to specify what kind of conclusion they would draw (if any) for each of the eight cases considered, and then to rank the cases according to their degree of confidence. They were asked to explain their reasoning aloud.

Results

Two types of attitudes were observed.

(1) The first, expressed by only three researchers, consisted in assuming that the conclusion was a problem which was largely outside the framework of the statistical analysis of data. The information gathered at the end of the analysis obviously entered into the picture, but it did not finish there; other information which did not pertain to the data, i.e. "outside" information such as references pertaining to theories, had to be explicitly taken into account and integrated into the reasoning which led to a conclusion. These researchers refused to propose a statistical conclusion before being able to link the result obtained to a certain "scientific consensus".

(2) The second, majority type of attitude, expressed by the other 20 researchers, consisted in drawing a conclusion based exclusively on the available statistical results (the t test and possibly the observed difference), that explicitly discarded *a priori* expectations. Their responses are summarized in Figure 2.

When the test was significant, even if all these researchers concluded in terms of the existence of a difference, it is important to note that nearly a third of them (6/20) tried to integrate the descriptive statistical results into their conclusion. They distinguished between the case in which the observed difference was small and the one in which it was large. Thus, when the observed difference was

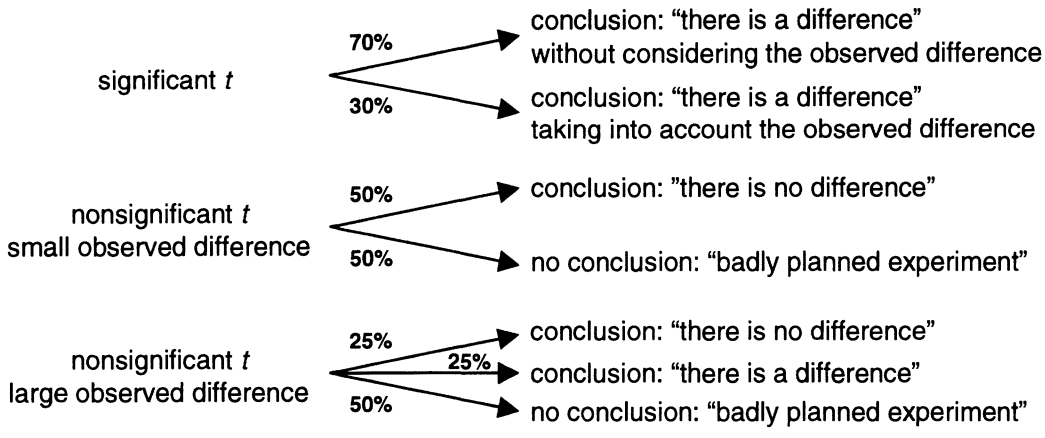

Figure 2. *Experiment 2: the conclusions given by the 20 majority researchers who discarded* a priori *information, according to the t test outcome and the observed difference.*

small, some researchers made qualified statements such as "there is probably a difference". Other researchers spoke in this case of a "small difference", whereas they spoke of a "strong effect" when the observed difference was large. Such statements resort to a kind of "naive" inference to integrate descriptive results and significance test results.

When the test was nonsignificant, the responses observed were very different amongst subjects. Firstly exactly half the subjects (10/20) drew a conclusion. When the observed difference was small (a non conflicting situation), all these ten subjects concluded improperly in terms of "no difference". When the observed difference was large (a conflicting situation), these subjects were divided and fell into two categories: five subjects relied on the nonsignificant test concluding, as in the previous case, that there was no difference, whereas the other five based themselves on the size of the observed difference concluding that there was a difference (however often with considerable reservations). The other ten subjects appeared very reluctant to draw a conclusion from a result which they perceived as "negative". These researchers declared that the result was either uninteresting or insufficient to pass judgment on, irrespective of the observed differences. Above all they tried to justify the result obtained by citing an error or an anomaly in the experimental conditions or the sample ("badly planned experiment"). They wanted to either repeat the experiment with a larger sample (in the hope of obtaining a significant result!) or to use some other statistical procedure.

Furthermore it is worth pointing out that the confidence ratings given by the researchers showed that they all had more confidence in their conclusion when the result was significant than when it was nonsignificant.

3.3 Experiment 3: Psychology Researchers and Professional Statisticians

20 psychology researchers and 25 professional statisticians were presented with the results of a study designed to test the efficacy of a drug by comparing two groups (treatment vs. placebo) with 15 patients in each. The following evaluation criterion of the efficacy of the drug was given to the subjects: the drug was considered as effective (clinically interesting) by experts in the field if the raw difference between the treatment and the placebo was more than +3. Four "result-situations" (see table 1) were constructed by combining the outcome of the *t* test (significant vs. nonsignificant) and

Table 1

The four result-situations and the corresponding normative answers based on the usual noninformative procedure for comparing two normal means with equal variances (see e.g., Box & Tiao, 1973; Lee, 1997).

Situation and appearance		t test (two-sided)	d	Posterior Bayesian probabilities $\delta < -3$	$\lvert\delta\rvert < 3$	$\delta > +3$	Conclusion
1	*non-conflicting*	+3.67 ($p = 0.001$)	+6.07	< 0.001	0.037	0.963	*effective*
2	*non-conflicting*	+0.68 ($p = 0.50$)	+1.52	0.026	0.719	0.256	*no firm conclusion*
3	*conflicting*	+3.67 ($p = 0.50$)	+1.52	< 0.001	0.999	0.001	*ineffective*
4	*conflicting*	+0.68 ($p = 0.001$)	+6.07	0.158	0.208	0.634	*no firm conclusion*

the observed difference between the two means d (large vs. small). Two of these situations appeared as conflicting (t significant/d small and t nonsignificant/d large).

From a normative viewpoint, situations 1 and 3 lead to the respective conclusions "effective" and "ineffective". On the contrary, situations 2 and 4 cannot lead to firm conclusions, because of the large variability observed. These normative answers can be legitimized by standard noninformative Bayesian statements (see Table 1), as well as by confidence intervals. The four situations were presented simultaneously. Three components of statistical inference were examined by means of open questions: drawing an inductive conclusion from the data in hand, making predictions for future data, and making a decision about stopping the experiment. With this purpose in mind, the following three questions were asked successively:

Question 1 – For each situation, what conclusion would you come to about the efficacy of the drug?

Question 2 – Initially the experiment was planned with 30 subjects in each group and the results presented here are in fact interim results. What would your prediction be for the final result, firstly for d then for t, and also for the conclusion about the efficacy of the drug?

Question 3 – From an economical viewpoint, it would obviously be interesting to stop the experiment with only the first 15 subjects in each group. For which of the four situations would you decide to stop the experiment and conclude?

It must be emphasized that all the subjects perceived the task as one they would frequently encounter in their profession. Only one statistician stated that he would have needed confidence intervals to reach the required conclusions.

Questions 1 (conclusion) and 3 (decision on stopping)

The responses given for Questions 1 and 3 are summarized in Figure 3.

Non-conflicting situations (1 and 2) gave rise to a large consensus about both the efficacy (*significant test, large d*) and the inefficacy (*nonsignificant test, small d*) of the drug. If this conclusion could be formally justified in Situation 1, Situation 2 was the case of a nonsignificant result incorrectly interpreted as a demonstration of "no effect" (or at least of a small effect). However Question 3 revealed less confidence in the conclusion for this situation: only a little more than half of the subjects who concluded that the drug was ineffective perceived the situation as being "very favorable" and then decided to stop the experiment (53% and 57% respectively for psychologists and statisticians).

Conflicting situations (3 and 4) exhibited differences between the two groups. In Situation 3 (*significant test, small d*), they were clearly divided on Question 1. Most statisticians (80%) concluded that the drug was not effective, correctly taking into account the smallness of d. On the contrary the psychologists were divided: almost half of them concluded that the drug was effective, relying exclusively on the result of the test and confusing "statistical significance" with "substantive significance". Situation 4 (*nonsignificant test, large d*) was considered as conflicting by a majority (65% of psychologists and 52% of statisticians) who did not give a conclusion. However it must be stressed that the test has such an impact that one-third of the subjects (35% and 36% respectively) erroneously

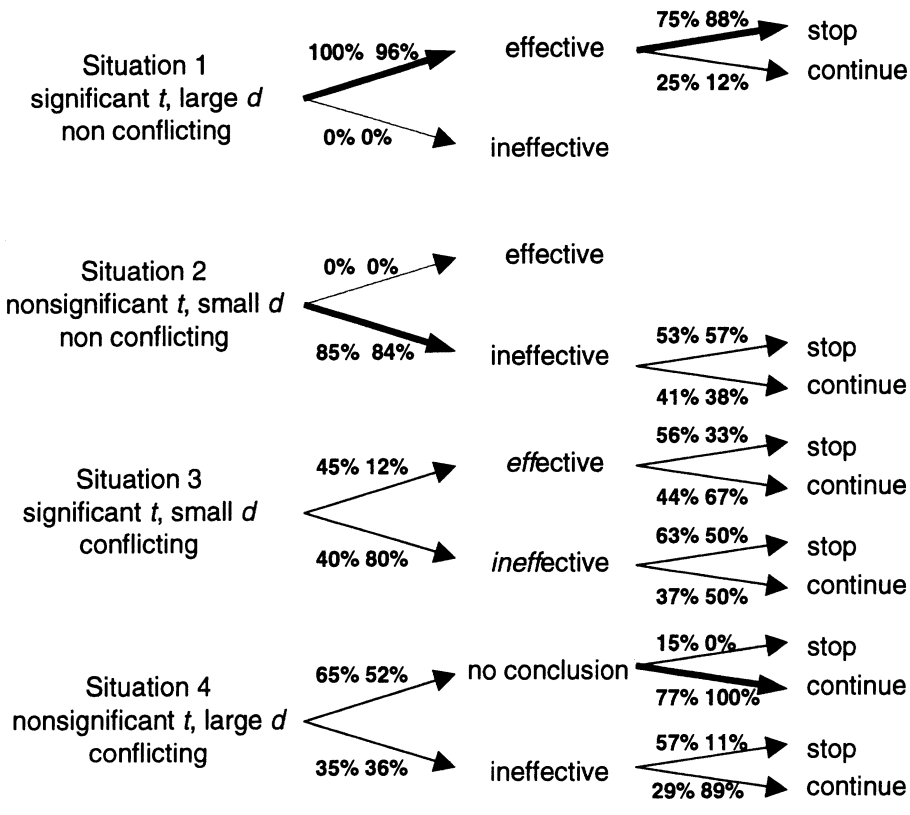

Figure 3. *Experiment 3: main responses (in percentage) given to Questions 1 and 3 for the four situations and the two groups. For each arrow, the first percentage refers to psychologists ($n = 20$) and the second one to statisticians ($n = 25$). Non reported minority responses are respectively "no conclusion" for Question 1, and "no decision" (situations 1, 4, 2) or "effective" (situation 3) for Question 3.*

concluded that the drug was ineffective, in spite of the large observed difference. Nevertheless in this case Question 3 distinguished these subjects: only one statistician (11%) as opposed to a majority of psychologists (57%), decided to stop the experiment, showing great confidence in their conclusion.

Question 2 (prediction)

In each of the four situations, when predicting the final result, the subjects in the two groups essentially answered either "about the same", which is generally the majority response, or "I can't predict anything". Most subjects did not differentiate their predictions about the significance test from those about the observed difference. Therefore predictive judgments available on the t test were generally incoherent, as the response "about the same" did not take into account the increase in sample size. This observation supports Freeman's (1993) conjecture that "even statisticians seem to have very little idea of how the interpretation of p-values should depend on sample size" (page 1446). It suggests that either our subjects had no conception of the t test statistic as an estimate of the experimental accuracy (conditionally on the observed difference), or used inappropriate heuristics such as the aforementioned representativeness hypothesis that could lead them to overlook the role of sample size.

On the whole psychologists and statisticians behaved in a similar way and were very impressed by statistically significant results. However psychologists were more reliant on the NHST outcome than statisticians were. Indeed when asking them to come to a conclusion about the efficacy of the drug almost half of them ignored the experts' criterion, above all in the case of significance but also in the case of nonsignificance. However it must be emphasized that professional statisticians were not immune from misinterpretations either, especially if the test was nonsignificant.

3.4 Discussion

The three experiments highlighted that only a minority of experienced users systematically had a dichotomous reject-accept attitude when spontaneously interpreting a null hypothesis significance test. On the contrary most of the subjects participating in the experiments rated graduated confidence judgments on p-values (Experiment 1), qualified the interpretation of significance tests by incorporating descriptive results (Experiments 2 and 3), or showed uncertainty in their conclusion when they were asked to decide whether the experiment should be stopped (Experiment 3). These findings must be opposed to the publication practice which dichotomizes each experimental result (significant vs. nonsignificant) according to the NHST outcome. This habit of treating NHST as a binary decision rule is undoubtedly encouraged by the decision-theoretic viewpoint often advocated in statistical literature. This is explicit within the Neyman–Pearsonian approach, but one can also consider, as Bakan (1966) did, that it is strongly suggested within the Fisherian approach. Furthermore this common practice probably reflects a circumstantial attitude ("it's the norm"), "mechanical behavior" (Gigerenzer, 1991), a socially approved "automatic routine" (Falk & Greenbaum, 1995). It is reinforced by a natural cognitive tendency to express clear-cut opinions in a publication and in some way arrange every NHST outcome in a "cognitive filing cabinet", in which a significant test is filed under "there is an effect" and a nonsignificant test is improperly filed under "there is no effect".

Further evidence for such a circumstantial attitude comes from the fact that only a few psychology researchers in Experiment 2 expressed arguments that acknowledge in a positive manner the role of NHST, such as "the test is like a gauge on the dashboard". In contrast more than three-quarters of the researchers expressed arguments that reflected a real consciousness of the stranglehold of NHST. In other words the significance test would only be used because "there is no other alternative". These subjects explicitly stated that they were dissatisfied with current practices. They expressed the need for inferential methods that would be better suited for answering their specific questions and would fit in better with their spontaneous interpretations of data. In this context a consensus consisted in expecting the statistical analysis to express in an objective way "what the data have to say" independently of any outside information. Indeed very few researchers stated that they wanted to integrate outside information—notably theoretical background—into the statistical analysis of data.

In any case a major finding obtained from all our experiments was the wide range of meanings that experienced users attach to null hypothesis significance tests. In actual fact the interpretation of tests could vary considerably from one individual to another, and it is hard to conceive that there could be a consensus. As a matter of fact it is not an easy task for experienced users to interpret p-values in a rational way. Particularly in the case of nonsignificance, most subjects appeared unable to combine the observed difference with the traditional t test properly (Experiment 3). If this can be interpreted as the inability to master NHST, this explanation can hardly be convincing for professional statisticians. At the very least whether NHST is able to meet the true needs of users can be seriously questioned. In fact beyond the superficial report of "erroneous" interpretations, one can see in the misuses of NHST intuitive judgmental "adjustments" (Bakan, 1966; Phillips, 1973, page 334), that try to overcome its inherent shortcomings. So the confusion between "statistical significance" and "substantive significance" illustrates such an adjustment, and can be seen as *adaptative abuse* designed to make an ill-suited tool fit the true needs of users. In the current context of the significant

test dictatorship, the nonsignificant case is again more illustrative. Faced with a nonsignificant result, users seem to have no other choice but to either interpret it as "proof of no effect" or attempt to justify it by citing an anomaly in the experimental conditions or in the sample. Note that this last attitude, frequently observed in Experiment 2, is the result of a general tendency to systematically try to obtain significant results. Many psychology researchers feel that any experiment that is "properly planned" should lead to statistically significant results. This is obviously a perverse research strategy that is adjusted to conform to the current normative context.

In conclusion experimental findings support our normative analysis that the use of null hypothesis significance tests is adapted to a social norm, but methodologically ill-suited.

4 Beyond the Significance Test Controversy: Prime Time for Bayes

4.1 Time for Change in Reporting Experimental Results

The times we're living in at the moment appear to be crucial. While users' uneasiness is ever growing (Lecoutre, 2000), changes in reporting experimental results, especially in presenting and interpreting effect sizes, are more and more enforced within editorial policies (see e.g., Rothman, 1978; Berry, 1986; Braitman, 1988, 1991; Loftus, 1993; Thompson, 1994, 1996; Heldref Foundation, 1997; Kendall, 1997; Murphy, 1997; Ellis, 2000; Hresko 2000; Kotrlick, 2000). Recent papers in psychology have developed concrete solutions (see e.g., Rogers, Howard & Vessey, 1993; Serlin & Lapsley, 1993; Loftus & Masson, 1994; Frick, 1995; Richardson, 1996; Rouanet, 1996; Schmidt, 1996; Brandstätter, 1999; Jones & Tukey; 2000; Lecoutre & Poitevineau, 2000). All these solutions are explicitly intended to deal with the question of *effect sizes*, which is essential "because science is inevitably about magnitudes" (Cohen, 1990, page 1309). Convergent proposals are constantly made in other fields, especially in medicine and pharmacology.

Reporting an effect size estimate is one of the first necessary steps in overcoming the abuses of NHST. It can effectively prevent researchers from unjustified conclusions in the conflicting cases where a nonsignificant result is associated with a large observed effect size. However our experiments reveal that small observed effect sizes are often illusorily perceived by researchers as being *favorable* to a conclusion of no effect, when they can't in themselves be considered as sufficient proof. Power studies can also be seen as a handrail to avoid hasty generalizations. However referring to statistical papers that discuss and compare procedures (for instance Schuirmann, 1987), a more and more widespread opinion is that "for interpretation of observed results, the concept of power has no place, and confidence intervals, likelihood, or Bayesian methods should be used instead" (Goodman & Berlin, 1994).

Nowadays the *official* trend is to advocate the use of confidence intervals, in addition or instead of NHST: see for instance the proposed guidelines for revising the statistical section of the American Psychological Association Publication Manual (Wilkinson *et al.*, 1999). So confidence intervals could quickly become a compulsory norm in experimental publications. Yet for many reasons due to their *frequentist* conception, confidence intervals can hardly be seen as the ultimate method.

4.2 Difficulties with Confidence Intervals

Indeed it can be anticipated that the conceptual difficulties encountered with the frequentist conception of confidence intervals will produce further dissatisfaction. In particular, users will realize that the appealing feature of confidence intervals is the result of a fundamental misunderstanding. As is the case with significance tests, the frequentist interpretation of a 95% confidence interval involves a long run repetition of the same experiment: in the long run 95% of computed confidence intervals will contain the "true value" of the parameter; each interval in isolation has either a 0 or 100% probability of containing it. Unfortunately treating the data as random even after observation

is so strange this "correct" interpretation does not make sense for most users. Ironically it is the interpretation in (Bayesian) terms of "a *fixed* interval having a 95% chance of including the true value of interest" which is the appealing feature of confidence intervals. Moreover these incorrect natural interpretations of confidence intervals (and of significance tests) are encouraged by most statistical teachers who tolerate and even use them: "... But it is perhaps of greater interest to be able to say whether or not some observed association in a sample of scores indicates that *the variables under study are most probably associated in the population* from which the sample was drawn." (Siegel, 1956, page 195, the italics are ours); "We can be 95% confident that the population mean is between 114.06 and 119.94." (Kirk, 1982, page 43). It can be emphasized with Rouanet (2000) that "it would not be scientifically sound to justify a procedure by frequentist arguments and to interpret it in Bayesian terms" (page 54). What a paradoxical situation! We then naturally have to ask ourselves whether the "Bayesian Choice" (Robert, 1994) will not, sooner or later, be unavoidable.

4.3 An Unjustified a Priori Against Bayesian Methods

Until now scientists have been reluctant to use Bayesian inferential procedures in practice. In a very lucid paper, which appears as if it had been written today, Winkler (1974) answered that "this state of affairs appears to be due to a combination of factors including philosophical conviction, tradition, statistical training, lack of availability, computational difficulties, reporting difficulties, and perceived resistance by journal editors" (page 129). If we leave to one side the choice of philosophical approach which is "not really as important as whether the approach is used consistently, carefully, and appropriately" (page 130), none of the aforementioned arguments are entirely convincing. However Bayesian methods are often felt to be too complicated to use and too subjective to be scientifically acceptable. The recent statement by Falk & Greenbaum (1995) illustrates this attitude: "Bayesian inference might, in principle, fill the void created by abandoning significance-testing", but "implementation of Bayesian analysis, however, requires subjective assessments of prior distributions, and often involves technical problems". Bayesians themselves are too often responsible for this mistrust. As Freeman (1993) puts it, "it is still wonder they are still treated as a kind of lunatic fringe preaching a doctrine so pure and untainted by the real world as to make it useful for little other than academics furthering their research careers" (page 1450). Consequently, without mentioning irrelevant caricature-like considerations (e.g., Chow, 1996), Bayesian methods for analysing experimental data have at best been constantly ignored, at worst discarded (e.g., Loftus & Masson, 1994; Frick, 1996) for *a priori* reasons that are more and more unjustified. Moreover the dominant frequentist conception, and the widespread use of significance tests, still appear to be such "a steamroller" (Berry, 1993) that even those who are open to the Bayesian approach often discard the Bayesian label so that their proposals are more likely to be accepted. For instance in a methodological paper for medical researchers, Goodman & Berlin (1994) give a very persuasive preliminary presentation of Bayesian methods. Yet having declared that "Bayesian posterior probabilities are exactly what scientists want", they then only discuss the use of confidence intervals, arguing that they are "more familiar" to readers than Bayesian probabilities.

4.4 The Bayesian Paradigm is Appropriate for Situations Involving Scientific Reporting

The contribution of Bayesian inference to experimental data analysis and scientific reporting has been obscured by the fact that many authors concentrate too much on the decision-theoretic elements of the Bayesian approach. "But the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested." (Rozeboom, 1960). Without dismissing the merits of the decision-theoretic viewpoint, it must be recognized that there is another approach which is just as

Bayesian which was developed by Jeffreys in the thirties. Following the lead of Laplace (1986/1825), this approach aimed at assigning the prior probability when nothing was known about the value of the parameter (see Jeffreys, 1961; Jaynes, 1983; without leaving out the precursory work by Lhoste, 1923). In practice, these *noninformative* prior probabilities are vague distributions which, *a priori*, do not favor any particular value. Consequently they let the data "speak for themselves" (Box & Tiao, 1973, page 2). In this form the Bayesian paradigm provides, if not objective methods, at least *reference* methods appropriate for situations involving scientific reporting.

In any case it must be acknowledged that any widely accepted inferential method cannot avoid more or less arbitrary conventions. For instance the arbitrariness of the choice of α in significance testing has been pointed out for a long time (e.g., Rozeboom, 1960; Camilleri, 1962; Winer, 1962, page 13; etc.). Neyman himself recognized an element of subjectivity in the theory of tests he founded with E. Pearson, for he firmly stated that the hypothesis to be tested (the so-called "null hypothesis", though not in Neyman's words) should be the one for which the risk of rejection if it is true must be controlled in priority, and this he admitted is a subjective matter (Neyman, 1950).

"At the very least, use of noninformative priors should be recognized as being at least as objective as any other statistical techniques." (Berger, 1985, page 110). The typical example of a Bernoulli process serves to illustrate this assertion. It is well known that in this case the NSHT procedure involves arbitrariness, especially in the specification of a stopping rule (Lindley & Phillips, 1976). Bernard (1996) demonstrated that the particular choice of a particular prior in an "ignorance zone" is *an exact counterpart* of the arbitrariness involved within the frequentist approach. This result can be generalized in the case of more than two categories (see for the case of a 2×2 contingency table Lecoutre & Charron, 2000). Another relevant reference is Walley (1996), where a notion of an imprecise Dirichlet model was developed.

4.5 Other Bayesian Techniques are Promising

If the use of noninformative priors has a privileged status in order to gain "public use" statements, other Bayesian techniques also have an important role to play in experimental investigations. They are ideally suited for combining information from several studies and therefore planning a series of experiments. Realistic uses of these techniques have been proposed. Various prior distributions expressing results from other experiments or subjective opinions from specific, well-informed individuals ("experts"), which whether *sceptical* or *enthusiastic*, can be investigated to assess the robustness of conclusions (see in particular Spiegelhalter, Freedman & Parmar, 1994). With regard to scientists' need for objectivity, it could be argued with Dickey (1986) that "an objective scientific report is a report of the whole prior-to-posterior mapping of a relevant range of prior probability distributions, keyed to meaningful uncertainty interpretations" (page 135).

In addition a major strength of the Bayesian paradigm is the ease with which one can make predictions about future observations. The predictive idea is central in experimental investigations, as "the essence of science is replication: a scientist should always be concerned about what would happen if he or another scientist were to repeat his experiment" (Guttman, 1983). Furthermore Bayesian predictive probabilities are effective tools for designing ("how many subjects?") and monitoring ("when to stop?") experiments (e.g., Choi & Pepple, 1989; Berry, 1991; Lecoutre, Derzko & Grouin, 1995; Dignam *et al.*, 1998; Johns & Andersen, 1999; Lecoutre, 2001). The predictive distribution of a test statistic can be used to include and extend the frequentist notion of power in a way that has been termed *predictive power* (Spiegelhalter, Freedman & Blackburn, 1986) or *expected power* (Brown *et al.*, 1987). More generally Bayesian predictive procedures give the researcher a very appealing method to evaluate the chances that the experiment will end up showing a conclusive result, or on the contrary a non-conclusive result. The prediction can be explicitly based on either the hypotheses used to design the experiment, expressed in terms of prior distribution, or on partially available data,

or on both.

4.6 The 21st Century: Reconciling Fisher and Bayes?

Curiously enough many critics and defenders of NHST who discuss its foundations ignore Fisher's conception of probability, which is of direct importance for the objectives Fisher assigned to statistical inference. Fisher firmly argued against the interpretation of the observed level as the relative frequency of error when sampling repeatedly from a same population (Fisher, 1990/1956, pages 81–82). His presentation of Student's t test explicitly did not refer to a frequentist conception (conditional on parameters), but on the contrary involved a predictive distribution conditional on the observed standard deviation (Lecoutre, 1985a). Like Bayesians, Fisher was evidently interested in inverse probability, as demonstrated not only by his work on the *fiducial* theory (e.g., Fisher, 1990/1956), but also by his work on the Bayesian method in his last years (Fisher, 1962). He was constantly concerned with considering a method that only expressed evidence from data in terms of probability about parameters and had *good conventional* properties. Fiducial inference is admittedly considered by most modern statisticians as a blunder, but it could be speculated with Efron (1998) that "maybe Fisher's biggest blunder will become a big hit in the 21st century" (page 107). We agree with him that "a widely accepted objective Bayes theory, which fiducial inference was intended to be, would be of immense theoretical and practical importance. A successful objective Bayes theory would have to provide good frequentist properties in familiar situations, for instance, reasonable coverage probabilities for whatever replaces confidence intervals." (page 106). In actual fact we suggest that such a theory is by no means a speculative viewpoint but on the contrary a desirable and perfectly feasible project.

4.7 The Fiducial Bayesian Methods

For many years we have worked with colleagues in France with this perspective in mind in order to develop standard "noninformative" Bayesian methods for the most familiar situations encountered in experimental data analysis (see in particular Rouanet, Lépine & Pelnard-Considère, 1976; Rouanet & Lecoutre, 1983; Lecoutre, 1985b; Lecoutre, Derzko & Grouin, 1995; Lecoutre, 1996; Rouanet, 1996; Bernard, 2000; Lecoutre & Charron, 2000; Lecoutre & Derzko, 2001; Lecoutre *et al.*, 2001).

In order to promote these Bayesian methods, it seemed important to us to give them a more explicit name than "standard", "noninformative" or "reference". We propose to call them *fiducial Bayesian*. This deliberately provocative name pays tribute to Fisher's work on scientific inference for research workers. It indicates their specificity and their aim to express "what the data have to say". These fiducial Bayesian methods are concrete proposals in order to bypass the shortcomings of NHST and improve current statistical methodology and practice (Rouanet *et al.*, 2000). Nowadays they are available and can be used as easily as the t, F or chi-square tests. Our statistical teaching and consulting experience, especially in psychology, showed us that they were far more intuitive and much closer to the thinking of scientists than frequentist procedures (Kadane, 1995). They have often been applied to real data and have been accepted well by psychology journals (see e.g., Hoc & Leplat, 1983; Ciancia *et al.*, 1988; Lecoutre, 1992; Hoc, 1996; Clément & Richard, 1997; and many experimental articles published in French).

5 Conclusion

"Null-hypothesis tests are not completely stupid, but Bayesian statistics are better." (Rindskopf, 1998). Based on more useful working definitions than frequentist procedures, Bayesian methods offer considerable flexibility, making all choices explicit. Bayesian routine procedures for familiar

situations in experimental data analysis are nowadays easy to implement and use. They offer promising new ways in statistical methodology. Their results can be presented in intuitively appealing and readily interpretable form. They provide scientists with relevant answers to essential questions raised by experimental data analysis.

However the use of NHST is such an integral part of scientists' behavior that its misuses and abuses should not be discontinued by flinging it out of the window. We suggest that the sole effective therapy for curing its "ills" is a *smooth transition* towards the Bayesian paradigm. Our strategy faced with the misuses of NHST is to introduce the Bayesian methods as follows. (1) To present natural Bayesian interpretations of NHST outcomes to draw attention to their shortcomings. (2) To create as a result of this the need for a change of emphasis in the presentation and interpretation of results. (3) Finally to equip users with a real possibility of thinking sensibly about statistical inference problems so that they behave in a more reasonable manner.

"We need statistical thinking, not rituals" (Gigerenzer, 1998). The Bayesian philosophy emphasizes the need to think hard about the information provided by the data in hand ("what do the data have to say?") instead of applying ready-made procedures. This should become an attractive challenge for scientists, applied statisticians and statistical teachers in the 21^{st} century.

Acknowledgments

We are especially grateful to Elja Arjas who encouraged the extension of an earlier version of this article and provided helpful suggestions and comments that have improved the final version. Our special thanks go to Victoria Bishop for improving our English. The remaining mistakes are ours.

References

A detailed bibliography is available upon request from the first author. The following web sites are dedicated to the "significance test controversy":
http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris.htm
http://acs.tamu.edu/~bbt6147/
http://www.cnr.colostate.edu/~anderson/null.html
http://www.indiana.edu/~stigsts/
http://www.npwrc.usgs.gov/perm/hypotest/

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, **66**, 423–437.
Beauchamp, K.L. & May, R.B. (1964). Replication report: interpretation of levels of significance by psychological researchers. *Psychological Reports*, **14**, 272.
Berger, G.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New-York: Springer Verlag.
Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, **33**, 526–542.
Bernard, J.-M. (1996). Bayesian interpretation of frequentist procedures for a Bernoulli process. *The American Statistician*, **50**, 7–13.
Bernard, J.-M. (2000). Bayesian inference for categorized data. In *New Ways in Statistical Methodology: From Significance Tests to Bayesian Inference* (2^{nd} edition), Eds. H. Rouanet, J.-M. Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre & B. Le Roux, pp. 159–226. Bern, CH: Peter Lang.
Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*. Chichester: John Wiley & Sons.
Berry, D.A. (1991). Experimental design for drug development: a Bayesian approach. *Journal of Biopharmaceutical Statistics*, **1**, 81–101.
Berry D.A. (1993). A case for Bayesianism in clinical trials. *Statistics in Medicine*, **12**, 1377–1393.
Berry, G. (1986). Statistical significance and confidence intervals [editorial]. *The Medical Journal of Australia*, **144**, 618–619.
Binder, A. (1963). Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, **70**, 107–115.
Bolles, R. (1962). The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports*, **11**, 639–645.
Boring, E.G. (1919). Mathematical versus scientific significance. *Psychological Bulletin*, **16**, 335–338.
Box, G.E.P. & Tiao, J.W. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison Wesley.
Braitman, L.E. (1988). Confidence intervals extract clinically useful information from data [editorial]. *Annals of Internal Medicine*, **108**, 296–298.

Braitman, L.E. (1991). Confidence intervals assess both clinical significance and statistical significance [editorial]. *Annals of Internal Medicine*, **114**, 515–517.

Brandstätter, E. (1999). Confidence intervals as an alternative to significance testing. *Methods in Psychological Research Online*, **4**, 33–46 (Internet: http://www. mpr-online.de).

Breslow, N. (1990). Biostatistics and Bayes [with comments]. *Statistical Science*, **5**, 269–298.

Brown, B.W., Herson, J., Atkinson, N. & Rozell, M.E. (1987). Projection from previous studies: a Bayesian and frequentist compromise. *Controlled Clinical Trials*, **8**, 29–44.

Camilleri, S.F. (1962). Theory, probability, and induction in social research. *American Sociological Review*, **27**, 170–178.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review. Harvard Educational Review*, **48**, 378–399.

Choi, S.C. & Pepple, P.A. (1989). Monitoring clinical trials based on predictive probability of significance. *Biometrics*, **45**, 317–323.

Chow, S.L. (1996). *Statistical Significance: Rationale, Validity and Utility*. London: Sage.

Ciancia, F., Maitte, M., Honoré, J., Lecoutre, B. & Coquery, J.-M. (1988). Orientation of attention and sensory gatting: an evoked potential and RT study in cat. *Experimental Neurology*, **100**, 274–287.

Clément, E. & Richard, J.-F. (1997). Knowledge of domain effects in problem representation: the case of Tower of Hanoi isomorphs. *Thinking and Reasoning*, **3**, 133–157.

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

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences* (2^{nd} edition; $1st$ edition: 1969). New-York: Academic Press.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, **45**, 1304–1312.

Craig, J.R., Eison, C.L. & Metze, L.P. (1976). Significance tests and their interpretation: An example utilizing published research and ω^2. *Bulletin of the Psychonomic Society*, **7**, 280–282.

Dickey J.M. (1986). Discussion of Racine, A., Grieve, A.P., Flühler, H. & Smith, A.F.M., Bayesian methods in practice: Experiences in the pharmaceutical industry. *Applied Statistics*, **35**, 93–150.

Dignam, J.J., Bryant, J., Wieand, H.S. *et al.* (1998). Early stopping of a clinical trial when there is evidence of no treatment benefit: protocol B-14 of the National Surgical Adjuvant Breast and Bowel Project. *Controlled Clinical Trials*, **19**, 575–588.

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

Efron, B. (1998). R.A. Fisher in the 21^{st} century [with discussion]. *Statistical Science*, **13**, 95–122.

Ellis, N. (2000). Editorial. *Language Learning*, **50**.

Evans, S.J.W., Mills, P. & Dawson, J. (1988). The end of the p-value? *British Heart Journal*, **60**, 177–180.

Falk, R. & Greenbaum, C.W. (1995). Significance tests die hard. The amazing persistence of a probabilistic misconception. *Theory & Psychology*, **5**, 75–98.

Fisher, R.A. (1962). Some examples of Bayes's method of the experimental determination of probabilities *a priori*. *Journal of the Royal Statistical Society, Series B*, **24**, 118–124.

Fisher, R.A. (1990/1925). *Statistical Methods for Research Workers*. London: Oliver and Boyd (14^{th} edition 1973 reprinted, Oxford University Press, 1990).

Fisher, R.A. (1990/1956). Oxford University Press, 1990). *Statistical Methods and Scientific Inference*. London: Oliver and Boyd (3^{rd} edition 1973 reprinted, Oxford University Press, 1990).

Fowler, R.L. (1985). Testing for substantive significance in applied research by specifying nonzero effect nullhypotheses. *Journal of Applied Statistics*, **70**, 215–218.

Freeman, P.R. (1993). The role of p-values in analysing trial results. *Statistics in Medicine*, **12**, 1443–1452.

Freiman, J.A., Chalmers, T.C., Smith, H. & Kueber, R.R. (1978). The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *New England Journal of Medicine*, **299**, 690–694.

Frick, R.W. (1995). Accepting the null hypothesis. *Memory & Cognition*, **23**, 132–138.

Frick, R.W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, **1**, 379–390.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: beyond "Heuristics and Biases". *European Review of Social Psychology*, **2**, 83–115.

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

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, **21**, 199–200.

Gigerenzer, G. & Murray, D.J. (1987). *Cognition as Intuitive Statistics*. Hillsdale, NJ: Erlbaum.

Glass, G.V., McGaw, B. & Smith, M.L. (1981). *Meta-analysis in Social Research*. Beverly Hills, CA: Sage.

Gold, D. (1969). Statistical tests and substantive significance. *The American Sociologist*, **4**, 42–46.

Goodman, S.N. & Berlin, J.A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, **121**, 200–206.

Grant, D.A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, **69**, 54–61.

Guttman, L. (1983). What is not what in statistics? *The Statistician*, **26**, 81–107.

Harcum, E.R. (1990). Methodological versus empirical literature: two views on casual acceptance of the null hypothesis. *American Psychologist*, **45**, 404–405.

Heldref Foundation (1997). Guidelines for contributors. *Journal of Experimental Education*, **65**, 95–96.

Hoc, J.-M. (1996). Operator expertise and verbal reports on temporal data. *Ergonomics*, **39**, 811–825.

Hoc, J.-M. & Leplat, J. (1983). Evaluation of different modalities of verbalization in a sorting task. *International Journal of Man-Machine Studies*, **18**, 283–306.

Hogben, L. (1957). *Statistical Theory: The Relationship of Probability, Credibility, and Error. An examination of the Contemporary crisis in Statistical Theory from a Behaviourist Viewpoint.* New-York: W.W. Norton.

Hresko, W. (2000). Editorial policy. *Journal of Learning Disabilities*, **33**, 214–215.

Jaynes, E. (1983). *Papers on Probability, Statistics, and Statistical Physics*. Ed. R.D. Rosenkrantz. Dordrecht, Netherlands: D. Reidel.

Jeffreys, H. (1961). *Theory of Probability* (3^{rd} edition; 1^{st} edition: 1939). Oxford: Clarendon.

Johns, D. & Andersen, J.S. (1999). Use of predictive probabilities in phase II and phase III clinical trials. *Journal of Biopharmaceutical Statistics*, **9**, 67–79.

Jones, L.V. & Tukey, J.W. (2000). A sensible formulation of the significance test. *Psychological Methods*, **5**, 411–414.

Kadane, J.B. (1995). Prime time for Bayes. *Controlled Clinical trials*, **16**, 313–318.

Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge.

Kahneman, D. & Tversky, A. (1972). Subjective probability: a judgement of representativeness. *Cognitive Psychology*, **3**, 430–454.

Kendall, P.C. (1997). Editorial. *Journal of Consulting and Clinical Psychology*, **65**, 3–5.

Kirk, R.E. (1982). *Experimental Design* (2^{nd} edition). Belmont: Brook-Cole.

Kish, L. (1959). Some statistical problems in research design. *American Sociological Review*, **24**, 328–338.

Kotrlick, J.W. (2000). Guidelines for authors. *Journal of Agricultural Education*, **41**, inside cover.

Laplace, P.-S. (1986/1825). *Essai Philosophique sur les Probabilités*. Paris: Christian Bourgois (English translation: *A Philosophical Essay on Probability*, 1952, New York: Dover).

Lecoutre, B. (1985a). Reconsideration of the F test of the analysis of variance: The semi-Bayesian significance tests. *Communications in Statistics-Theory and Methods*, **14**, 2437–2446.

Lecoutre, B. (1985b). How to derive Bayes-fiducial conclusions from usual significance tests. *Cahiers de Psychologie Cognitive*, **5**, 553–563.

Lecoutre, B. (1996). *Traitement Statistique des Données Expérimentales: Des Pratiques Traditionnelles aux Pratiques Bayésiennes* (with Bayesian Windows programs by B. Lecoutre and J. Poitevineau, freely available upon request: mail to bruno.lecoutre@univ-rouen.fr). Montreuil (France): CISIA.

Lecoutre, B. (2001). Bayesian predictive procedure for designing and monitoring experiments. ISBA 2000 and Eurostat: *ISBA 2000, Proceedings*, in press.

Lecoutre, B. & Charron, C. (2000). Bayesian procedures for prediction analysis of implication hypotheses in 2×2 contingency tables. *Journal of Educational and Behavioral Statistics*, **25**, 185–201.

Lecoutre, B. & Derzko, G. (2001). Asserting the smallness of effects in ANOVA. *Methods of Psychological Research*, **6**:1, 1–32. [http://www.mpr-online.de/].

Lecoutre, B., Derzko, G. & Grouin, J.-M. (1995). Bayesian predictive approach for inference about proportions. *Statistics in Medicine*, **14**, 1057–1063.

Lecoutre, B., Lecoutre, M.-P. & Grouin, J.-M. (2001). A challenge for statistical instructors: Teaching Bayesian inference without discarding the "official" significance tests. ISBA 2000 and Eurostat: *ISBA 2000, Proceedings*, in press.

Lecoutre, B. & Poitevineau, J. (2000). Aller au delà des tests de signification traditionnels: vers de nouvelles normes de publication. *L'Année Psychologique*, **100**, 683–713.

Lecoutre, M.-P. (1982). Comportement des chercheurs dans des situations conflictuelles d'analyse des données expérimentales. *Psychologie Française*, **27**, 1–8.

Lecoutre, M.-P. (1983). La démarche du chercheur en psychologie dans des situations d'analyse statistique de données expérimentales. *Journal de Psychologie Normale et Pathologique*, **3**, 275–295.

Lecoutre, M.-P. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics*, **23**, 557–568.

Lecoutre, M.-P. (2000). And ... What about the researcher's point of view. In *New Ways in Statistical Methodology: From Significance Tests to Bayesian Inference* (2^{nd} edition), H. Rouanet, J.-M. Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre & B. Le Roux, pp. 65–95. Bern, CH: Peter Lang.

Lecoutre, M.-P. & Rouanet, H. (1993). Predictive judgements in situations of statistical analysis. *Organizational Behavior and Human Decision Processes*, **54**, 45–56.

Lee, P. (1997). *Bayesian Statistics: An Introduction* (2^{nd} edition), London: Arnold.

Lhoste, E. (1923). Le calcul des probabilités appliqué à l'artillerie. *Revue d'Artillerie*, **91**.

Lindley, D.V. & Phillips, L.D. (1976). Inference for a Bernoulli process (a Bayesian view). *The American Statistician*, **30**, 112–119.

Loftus, G.R. (1993). Editorial comment. *Memory & Cognition*, **21**, 1–3.

Loftus, G.R. & Masson, M.E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, **1**, 476–490.

McNemar, Q. (1960). At random: sense and nonsense. *American Psychologist*, **15**, 295–300.

Mittag, K.C. & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher*, **29**, 14–20.

Morrison, D.E. & Henkel, R.E. (1969). Significance tests reconsidered. *The American Sociologist*, **4**, 131–140.

Murphy, K.R. (1997). Editorial. *Journal of Applied Psychology*, **82**, 3–5.

Natrella, M.G. (1960). The relation between confidence intervals and tests of significance. *The American Statistician*, **14**, 20–22.

Nelson, N., Rosenthal, R. & Rosnow, R.L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*, **41**, 1299–1301.

Neyman, J. (1950). *First Course in Probability and Statistics*. New York: Holt.

Neyman, J., Scott, E.L. & Smith, J.A. (1969). Letter in *Science*, **165**, 618.

Nisbett, R. & Ross, L. (1981). *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice Hall: Century Psychology Series.

Nunnally, J.C. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, **20**, 641–650.

Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioural Sciences*. New-York: Wiley.

O'Brien, T.C. & Shapiro, B.J. (1968). Statistical significance - What? *Mathematics Teacher*, **61**, 673–676.

Phillips, L.D. (1973). *Bayesian Statistics for Social Scientists*. London: Nelson.

Poitevineau, J. (1997). *Méthodologie de l'Analyse des Données Expérimentales: Etude de la Pratique des Tests Statistiques chez les Chercheurs en Psychologie*. Doctoral Thesis, Université de Rouen (France).

Reuchlin, M. (1962). *Les Méthodes Quantitatives en Psychologie*. Paris: Presses Universitaires de France.

Richardson J.T.E. (1996). Measures of effect size. *Behavior Research Methods, Instruments & Computers*, **28**, 12–22.

Rindskopf, D.(1998). Null-hypothesis tests are not completely stupid, but Bayesian statistics are better. *Behavioral and Brain Sciences*, **21**, 215–216.

Robert, C. (1994). *The Bayesian Choice: A Decision-theoretic Approach*. New York: Springer Verlag.

Rogers, J.L., Howard, K.I. & Vessey, J. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, **113**, 553–565.

Rosenthal, R. & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, **55**, 33–38.

Rosenthal, R. & Gaito, J. (1964). Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports*, **15**, 570.

Rothman, K.J. (1978). A show of confidence [editorial]. *The New England Journal of Medicine*, **299**, 1362–1363.

Rouanet, H. (1996). Bayesian procedures for assessing importance of effects. *Psychological Bulletin*, **119**, 149–158.

Rouanet, H. (2000). Statistical Practice revisited. In *New Ways in Statistical Methodology: From Significance Tests to Bayesian Inference* (2^{nd} edition), Eds. H. Rouanet, J.-M. Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre & B. Le Roux, pp. 29–64. Bern, CH: Peter Lang.

Rouanet, H., Bernard, J.-M., Bert, M.-C., Lecoutre, B., Lecoutre, M.-P. & Le Roux, B. (2000). *New Ways in Statistical Methodology: From Significance Tests to Bayesian Inference* (2nd edition). Bern, CH: Peter Lang.

Rouanet, H. & Lecoutre, B. (1983). Specific inference in ANOVA: from significance tests to Bayesian procedures. *British Journal of Mathematical and Statistical Psychology*, **36**, 252–268.

Rouanet, H., Lépine, D. & Pelnard-Considère, J. (1976). Bayes-Fiducial procedures as practical substitutes for misplaced significance testing: an application to educational data. In *Advances in Psychological and Educational Measurement*, Eds. D.N.M. De Gruijter D.N.M. & L.J.T. Van Der Kamp, pp. 33–50. New-York: Wiley.

Rozeboom, W.W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, **57**, 416–428.

Salsburg, D.S. (1985) The religion of statistics as practiced in medical journals. *The American Statistician*, **39**, 220–223.

Schervish, M.J. (1995). *Theory of Statistics*. New York: Springer Verlag.

Scheutz, F., Andersen, B. & Wulff, H.R. (1988). What do dentists know about statistics? *Scandinavian Journal of Dental Research*, **96**, 281–287.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychological Methods*, **1**, 115–129.

Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, **15**, 657–680.

Selvin, H.C. (1957). A Critique of tests of significance in survey research. *American Sociological Review*, **22**, 519–527.

Serlin, R.C. & Lapsley, D.K. (1993). Rational appraisal of psychological research and the good-enough principle. In *A Handbook for Data Analysis in the Behavioral Sciences. Vol 1: Methodological Issues*, Eds. G. Keren & C. Lewis, pp. 199–228. Hillsdale, N-J: Erlbaum.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

Smith, A. (1995). A conversation with Dennis Lindley. *Statistical Science*, **10**, 305–319.

Spiegelhalter, D.J., Freedman, L.S. & Blackburn, P.R. (1986). Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials*, **7**, 8–17.

Spiegelhalter, D.J., Freedman, L.S. & Parmar, M.K.B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society A*, **157**, 357–416.

Sterling, T.D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, **54**, 30–34.

Sterling, T.D., Rosenbaum, W.L. & Weinkam, J.J. (1995). Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, **49**, 108–112.

Student (1908). The probable error of a mean. *Biometrika*, **6**, 1–25.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, **54**, 837–847.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: three suggested reforms. *Educational Researcher*, **25**, 26–30.

Tullock, G. (1959). Publication decisions and tests of significance: a comment. *Journal of the American Statistical Association*, **54**, 593.

Tversky, A. & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, **76**, 237–251.

Tyler, R. (1931). What is statistical significance? *Educational Research Bulletin*, **10**, 118–142.

Victor, N. (1987). On clinically relevant differences and shifted nullhypotheses. *Methods of Information in Medicine*, **26**, 109–116.

Walley, P. (1996). Inferences from multinomial data: learning about a bag of marbles [with discussion]. *Journal of the Royal Statistical Society B*, **58**, 3–57.

Wilkinson, L. and Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist*, **54**, 594–604.

Winch, R.F. & Campbell, D.T. (1969). Proof? No. Evidence? Yes. The significance of tests of significance. *The American Sociologist*, **4**, 140–143.

Winer, B. J. (1962). *Statistical Principles in Experimental Designs*. New York: McGraw-Hill.

Winkler, R.L. (1974). Statistical analysis: theory versus practice. In *The Concept of Probability in Psychological Experiments*, Ed. C.-A.S. Staël Von Holstein, pp. 127–140. Dordrecht, Holland: D. Reidel.

Wulff, H.R., Andersen, B., Brandenhoff, P. & Guttler, F. (1987). What do doctors know about statistics? *Statistics in Medicine*, **6**, 3–10.

Zuckerman, M., Hodgins, H., Zuckerman, A. & Rosenthal, R. (1993). Contemporary issues in the analysis of data: a survey of 551 psychologists. *Psychological Science*, **4**, 49–53.

Résumé

Nous discutons d'abord brièvement le contexte actuel de la "controverse sur le test de signification". Puis nous présentons des recherches expérimentales sur l'usage des tests de signification de l'hypothèse nulle par des chercheurs scientifiques et des statisticiens professionnels. Les mauvais usages de ces tests sont reconsidérés comme des jugements adaptatifs, qui révèlent les exigences des chercheurs envers l'inférence statistique. Finalement nous envisageons les solutions de rechange. Nous en venons naturellement à poser la question: "le choix bayésien ne sera-t-il pas incontournable?"

[*Received December 1999, accepted August 2000*]