JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

DANIEL J. DENIS

The modern hypothesis testing hybrid : R. A. Fisher's fading influence

Journal de la société française de statistique, tome 145, n° 4 (2004), p. 5-26

<http://www.numdam.org/item?id=JSFS_2004__145_4_5_0>

THE MODERN HYPOTHESIS TESTING HYBRID: R. A. FISHER'S FADING INFLUENCE

Daniel J. DENIS *

ABSTRACT

Today's genre of null hypothesis significance testing (NHST) bears little resemblance to the model originally proposed by Fisher over seventy-five years ago. Aside from general misunderstandings, the present model incorporates features that Fisher adamantly rejected. The aim of this article is to bring to attention how NHST differs from the model first proposed by Fisher in 1925, and in doing, locate his model within today's hybrid of hypothesis testing. It is argued that associating Fisher's name with today's version of NHST is not only incorrect, it inappropriately blames Fisher for NHST's deep methodological and philosophical problems. An attempt is made to distinguish between Fisher's original model and today's hybridized, and generally misunderstood approach to statistical inference. It will be shown that today's social science researchers utilize a logically faulty and distasteful blend of Fisherian, Neyman-Pearson and Bayesian ingredients.

RÉSUMÉ

De nos jours la nature du « test de signification d'une hypothèse nulle » (NHST) présente peu de ressemblance avec le modèle proposé par Fisher il y a quelque quatre-vingts ans. Au-delà de certains malentendus, le modèle actuel incorpore des aspects que Fisher rejetait fermement. Le but de cet article est de mettre en évidence la façon dont le NHST diffère du modèle proposé par Fisher en 1925 et, ce faisant, de resituer le modèle initial par rapport aux méthodes hybrides actuelles. On montre qu'associer le nom de Fisher à ces dernières non seulement est incorrect, mais encore adresse à Fisher des reproches injustifiés au sujet des profondes faiblesses méthodologiques et philosophiques du NHST. On essaie de distinguer entre la méthodologie originale de Fisher et l'hybride actuelle, et une approche généralement mal comprise de l'inférence statistique. On montre que les chercheurs en sciences sociales utilisent aujourd'hui un défectueux et déplaisant mélange des ingrédients dus à Fisher, Neyman-Pearson et Bayes.

* York University, now at Department of Psychology, University of Montana, Missoula, Montana 59812, Canada.
A shorter version of this article was presented at the 107th Annual Convention of the American Psychological Association, Boston, MA. The author is indebted to Christopher D. Green who provided invaluable feedback and suggestions on previous drafts of this article. Bruno Lecoutre was also of great assistance and his help was much appreciated.
e-mail : daniel.denis@umontana.edu.

Null hypothesis significance testing (NHST) as practiced by today's community of social scientists suffers from deep theoretical and philosophical insufficiencies (*e.g.*, see Bakan, 1966; Carver, 1993; Cohen, 1990, 1994). It has historically been a favorite target of criticism by methodologists since its original inception by R. A. Fisher in 1925. As noted by Gigerenzer (1993), today's model of hypothesis testing is best considered a "hybrid" of Fisherian, Neyman-Pearson and Bayesian approaches. In the present piece, I provide an historical overview of null hypothesis significance testing (NHST) [1] focusing primarily on Fisher. The components of Fisher's model are drawn out in detail for the purpose of staging a contrast and comparison between his *original* model and later modifications that were added to this early configuration. The Neyman-Pearson and Bayesian approaches to hypothesis testing are discussed as partial "contributors" to today's hybridized model. Through an evaluation of how today's model incorporates little of what Fisher originally prescribed, the objective is to show how today's misused and misunderstood model should hardly at all be considered Fisherian. A comparison between today's model, early and late Fisherian models, the Neyman-Pearson model, and the Bayesian approach is provided in evaluating the claim that today's null hypothesis significance testing is attributable to Fisher. The following will show that although many scientists use a *similar* model to that once proposed by Fisher, today's researchers use anything but a *pure* Fisherian approach. Despite this hybridization of hypothesis testing procedures, today's NHST is still commonly regarded by social scientists as "Fisherian" (Cowles, 1989). As a result of such misattribution, Fisher has been on occasion unjustly denounced for problems associated with today's model, a model that he did not advocate. An instance of such misguided criticism will be given along with a typical empirical example that highlights the hybridized interpretation of NHST. As will become apparent, Fisher's influence is forever fading from modern hypothesis testing procedures. Cowles, also aware of today's pseudo-Fisherian model, made a comment that so aptly describes the thesis I defend:

> Perhaps we should spare a thought for Sir Ronald Fisher, curmudgeon that he was. He must indeed be constantly tossing in his grave as lecturers and professors across the world, if they remember him at all, refer to the content of most current curricula as *Fisherian statistics.* (Cowles, 1989, p. 189)

1. Null Hypothesis Significance Testing: Fisher's Original Paradigm

Before diving into Fisher's significance testing principles, it is perhaps wise to first comment on Fisher's view of induction and inference in the context of experimental design. In *Design of Experiments* (1966), Fisher's introductory chapter delineates his views regarding mathematical induction. He argued

1. Although the arguments presented in this article are taken primarily from the field of psychology, similar arguments are applicable to allied fields such as sociology and biology. The philosophical problems associated with NHST are relatively constant across various fields of application.

for the estimation of population parameters based on small sample data. Although results may be probabilistic, Fisher saw no problem with this:

Many mathematicians, if pressed on the point, would say that it is not possible rigorously to argue from the particular to the general; that all such arguments must involve some sort of guesswork, which they might admit to be plausible guesswork, but the rationale of which, they would be unwilling, as mathematicians, to discuss. We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression... The mere fact that inductive inferences are uncertain cannot, therefore, be accepted as precluding perfectly rigorous and unequivocal inference. (Fisher, 1966, pp. 3-4)

Hence, Fisher argued for the *rigorous* quantification of uncertainty when drawing inferences from samples to populations. He believed that scientific inference can be exact, even if uncertain (Fisher, 1935b). In other words, an uncertain (*i.e.*, probabilistic) inference can be as precise as one that is certain. Fisher's philosophy of science held that we learn from experience, yet knowledge must always remain provisional. Knowledge is uncertain, but this uncertainty can be quantified using appropriate statistical measures (Gigerenzer, Swijtink, Porter, Daston, Beatty, Krüger, 1989). It is the development of such measures that would occupy Fisher throughout much of his productive life.

1.1. Forecasting of Results

In considering now the components of Fisherian significance testing, it is appropriate to begin with a basic prescription made by Fisher; that of always forecasting beforehand all possible results of the experiment. Furthermore, he asserted that we must know in advance the interpretation of each of the given possibilities. Fisher stated:

It is always needful to forecast all possible results of the experiment, and to have decided without ambiguity what interpretation shall be placed upon each one of them. Further, we must know by what argument this interpretation is to be sustained. (Fisher, 1966, p. 12)

Fisher required that the experimenter know *in advance* the possible outcomes of the given experiment. This would require the experimenter to calculate the probability of a given result occurring by chance alone. This is typically accomplished using mathematical permutations and combinations. It is a relatively straightforward task. For example, a correct hand grab from a subject claiming to be able to "psychically" select the marked ball from an urn containing a total of just two balls would not impress in the least, since most would agree that the probability of selecting the marked ball is 0.5. On the other hand, should there be a total of 1000 balls in the urn and the subject successfully selects the correct ball, this may be deemed justification in rejecting the hypothesis that the selection was better explained by chance. This outcome is presumably more likely to be used as a rationale for refuting

the chance hypothesis because the probability (at least in frequentist terms) of selecting the one ball from a total of 1000 balls is equal to 0.001 (or 1 in 1000), making it an extremely unlikely event.

Fisher's second requirement noted above, although somewhat more difficult to satisfy, is just as important. In arguing that the experimenter must know in advance the interpretation of each possible outcome, Fisher placed great weight and value on fully describing the *design* of the experiment *before* the data are collected and analyzed. Using our example, this requirement would have the experimenter state in advance his or her interpretation of possible results before the subject reaches in the urn to choose a ball. For Fisher then, the experimenter must have adequately designed the experiment before the data are collected, and for Fisher, "design" included the anticipation of possible outcomes, along with their respective interpretations. After the data are collected, there should be no surprises. Later in his career however, Fisher (1956) did recommend that the *exact* significance level be reported *after* the analysis of the data. Hence, this contradicted somewhat his earlier recommendation (Fisher, 1935a) that the significance level be determined before the experiment is executed. A more thorough treatment and discussion of significance levels is given later in this article.

1.2. Randomization

The second key component of Fisherian significance testing is that of randomization (Fisher, 1925, 1966). Fisher was adamant with regards to the randomization of subjects to treatments if an experiment were to be considered at all meaningful. His recommendations for randomization were strict, with each subject having to be randomly assigned to each experimental condition. The assignment of subjects to conditions would likely be different had the experimenter allocated them, and hence not be random, for she might subconsciously let her opinions influence the allocation of subjects (Gigerenzer *et al.*, 1989). Although in some cases error could actually be reduced by systematic allocation ("Student", 1937), Fisher was more concerned with the validity of the *estimates* of error, than the quantity of error.

For Fisher, randomization was necessary to satisfy the assumption that should the null hypothesis fail to be rejected, the experimental result was better explained as being credited to chance or sampling error. Although randomization did not eliminate all possible sources of bias in the experiment, it did minimize potential error. As Fisher explained:

Apart, therefore, from the avoidable error of the experimenter himself introducing with his test treatments, or subsequently, other differences in treatment, the effects of which the experimenter is not intended to study, it may be said that the simple precaution of randomisation will suffice to guarantee the validity of the test of significance, by which the result of the experiment is to be judged. (Fisher, 1966, p. 21).

1.3. Infinite Hypothetical Populations

A third component of Fisher's statistical theory was that the population could not be *known* per se. That is, when a sample is drawn, it is impossible for the researcher to have specified beforehand the population from which the sample was chosen. Rather, the population was hypothetical. Fisher argued:

There is always, as Venn (1876) in particular has shown, a multiplicity of populations to each of which we can legitimately regard our sample as belonging; so that the phrase "repeated sampling from the same population" does not enable us to determine which population is to be used to define the probability level, for no one of them has objective reality, all being products of the statistician's imagination. (Fisher, 1955, p. 71).

Exactly what Fisher meant by "infinite hypothetical population" is not at all clear. Kendall, obviously confused by Fisher's claim, stated, "This is, to me at all events, a most baffling conception" (Kendall, 1943, p. 17). Gigerenzer has noted that "the concept of an unknown hypothetical infinite population has puzzled many" (Gigerenzer, 1993, p. 321). Indeed, the logic behind Fisher's argument has been questioned by some. Opposition to Fisher's controversial construct will be discussed later in this article.

1.4. Testing of the Null Hypothesis

A fourth component of Fisher's statistical theory was the testing of just one hypothesis – the null hypothesis. The duality of a null *versus* an alternative hypothesis was introduced by Neyman and Pearson (1928), and formed an integral part of their model of hypothesis testing in the context of decision making. Fisher was adamantly against any implication of testing or a commitment to choose an alternative hypothesis to account for experimental results not explained by the null. As Fisher explained:

It might be argued that if an experiment can disprove the hypothesis that the subject possesses no sensory discrimination between two different sorts of object, it must therefore be able to prove the opposite hypothesis, that she can make some discrimination. But this last hypothesis, however reasonable or true it may be, is ineligible as a null hypothesis to be tested by experiment, because it is inexact. (Fisher, 1966, p. 16).

In this, Fisher held that the opposite hypothesis, or alternative hypothesis, can never be staged as a hypothesis to be "nullified" because it is not precise enough to be under test. The null hypothesis, as Fisher explained, "must be exact, that is free from vagueness and ambiguity, because it must supply the basis of the 'problem of distribution', of which the test of significance is the solution" (Fisher, 1966, p. 16). Since the alternative hypothesis does not exhibit these characteristics, it is invalid to test it in any way with a significance test, and it is questionable whether one can infer it when the null is shown to be false. Regarding the treatment of the null, Fisher wrote:

It should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment

may be said to exist only in order to give the facts a chance of disproving the null hypothesis. (Fisher, 1966, p. 16).

Therefore, no matter how many times a null fails to be rejected, it never in itself is proved. A null hypothesis can *never* be shown to be true. All the experimenter can hope for is to possibly reject the null hypothesis. That for Fisher was the purpose of using significance testing in an experiment. As noted in Gigerenzer (1993), Fisher later said that, "It is a fallacy... to conclude from a test of significance that the null hypothesis is thereby established; at most it may be said to be confirmed or strengthened" (Fisher, 1966, p. 73). From this it would appear that Fisher was leaning towards a confirmation theory of the null, yet this inference depends on how one interprets his use of the term "established" as being different from the term "confirmed". As Gigerenzer (1993) noted, Fisher never explained further how a non-significant result might possibly act as support for the null hypothesis. The reader is left somewhat confused by Fisher's writings.

1.5. Significance Levels

A fifth component of Fisherian statistics is that of significance levels. Fisher was vague as to what level of significance the researcher should adopt in testing the null hypothesis. This ambiguity is hardly a surprising feature of his work on significance testing. As Gigerenzer noted, his writings on significance testing "had a remarkably elusive quality, and people have read his work quite differently" (Gigerenzer, 1993, p. 316). His recommendations were often conflicting. For instance, early in *Design of Experiments*, Fisher wrote, "It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result" (Fisher, 1966, p. 13).

Later however, on the same page, Fisher wrote the following:

It is usual and convenient for experimenters to take 5 per cent. as a standard level of significance, in the sense that they are prepared to *ignore* [emphasis added] all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results. (Fisher, 1966, p. 13).

Fisher also argued, much in response to the "alpha" definition proposed by Neyman and Pearson (1928), that "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" (Fisher, 1956, p. 42). Through these passages, Fisher gave an ambiguous instruction as to which significance levels to use and when to use them. It should be emphasized however that Fisher's most extreme recommendation for probability values at or below the .05 level, was that "it is usual and convenient"; he never implied that a paper's scientific value should be judged on this basis alone, or that publication decisions should be made on meeting this sole criterion. Indeed, as noted by his daughter, Joan Fisher Box, later in his career, Fisher himself regarded the significance test to

be a rather "weak" argument. As Fisher Box commented: "much" of his early work [Fisher's work] has been devoted to what he came to regard as the lowest level of scientific inference – to tests of significance which make a dichotomy between hypotheses that are discredited by the data and those that are not" (Fisher Box, 1978, pp. 447-448).

In summarizing Fisher's notion of significance levels, Gigerenzer (1993) argued for two categories for his ideas. The first is that of a *standard level of significance*, which consists of a conventional standard (*i.e.*, 0.05) that could be adopted by researchers. This was Fisher's early position. The second position became apparent near the end of Fisher's career; that of an *exact level of significance*, for which the level (the exact level, *e.g.*, 0.03) was noted in publication. It would appear that researchers adopted Fisher's early view despite what he had to say later in his career. The concept of significance levels remains perhaps the most important feature of Fisherian significance testing. Yet because of Fisher's ambiguity in explaining this all-important concept, they remain quite possibly the most misunderstood and controversial component of his entire statistical theory. However, to evaluate all research results on a rigid and dogmatic criterion such as $p < 0.05$ is to restrict one's interpretation of science wholly to statistical arithmetic at the expense of a balanced view of a research paradigm. As Boring succinctly stated, "statistical ability, divorced from a scientific intimacy with the fundamental observations, leads nowhere" (Boring, 1919, p. 338).

1.6. Publish Positive and Negative Results

Related to significance levels were Fisher's ideas regarding publication policies. According to Gigerenzer *et al.* (1989), Fisher's discussion of the relation between a significant result and the demonstration of a phenomenon suggests that both significant and non-significant results should be published, for the purpose of being able to compare the relative frequency of the significant to the non-significant results. This in turn would supply the literature with a relative comparison and, through this, the establishment of a phenomenon would become apparent. However, a precise ratio of "significant *vs.* non-significant" results that would serve to demonstrate the existence of a phenomenon was never outlined by Fisher. More recently, the problem of not accounting for non-significant results has been called the "file drawer problem". As noted by Rosenthal (1979), the problem arises when one considers the possibility that journals are filled with the 5% of studies that constitute Type I errors, while those studies not published (*i.e.*, the file drawers) are filled with the 95% that show non-significant results. Had significance testing remained Fisherian, the file drawer problem would likely not exist today.

This last component may be argued to have little to do with Fisher's theory per se, and everything to do with publication policy. I would venture to disagree with this and hold that because significance levels are so influential in publication decisions, publication should be discussed as part of Fisherian theory. If publication is to include those documents that are part of "knowledge" in general, then what is allowed to be included in that category has serious

implications for what is acknowledged as progress in a field of study. In other words, publication is a derivative of the word "public", and it is assumed that anything not published is for all purposes not known to the community of researchers. Having said this, I quote Fisher:

"In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment that will rarely fail to give us a statistically significant result" (Fisher, 1966, p. 14).

In this, Fisher implied that both significant and non-significant results should be published. *Fisher's words are of extreme importance.* How else can we account for both positive and negative results if they are not published ? How are we supposed to know how many "failures" occur if we do not document them, as we do positive results ? Fisher would have it that a ratio of positive publications be contrasted with negative publications that would in turn represent the existence or non-existence of a phenomenon. However, as already mentioned, he did not specify the magnitude this ratio should take before a phenomenon is reputed as "existing". There is no question that Fisher would not regard a single significant result as evidence for the existence of a phenomenon, but we are unfortunately left with an incomplete account of how Fisher would address various ratios. This naturally leads to the question of what ratio of significant to non-significant results would Fisher accept as deeming a phenomenon "significant" ? Unfortunately, he provided us with no answer except to say that any result is provisional upon further experimentation.

1.7. Sensitivity of Experiments

A seventh feature of Fisherian significance testing concerns the *sensitivity* of an experiment. What Neyman and Pearson (1928) called *power* can be closely allied, at least in a conceptual sense, to Fisher's sensitivity. This last claim is made with some reservation since Fisher never used, nor liked the term "power". Further, we cannot fully equate Fisherian sensitivity with Neyman and Pearson power since the latter is a conditional probability closely related to Type II errors. Of course, Fisher's significance testing theory had little tolerance for the possibility of Type II errors. What ties Fisherian sensitivity and Neyman Pearson power together is that both are intimately related to sample size. Fisher recognized the usefulness of considering sample size in relation to desired *effect size.* Fisher also noted that one can increase the sensitivity of an experiment by ensuring proper controls (*i.e.*, controlling potential covariates) and hence reducing error variability. Further elaboration on these points had to wait until Neyman and Pearson properly defined power. Cohen (1962) later contributed enormously to the concept of power by providing relatively easy computational methods. What is important to note is that Fisher did acknowledge the importance of sample size (which is the major determinant of statistical power) and estimating effect size when designing the "model experiment". Although Fisher rejected the concept of power as propounded by Neyman and Pearson, I argue that while he may

have disregarded the way Neyman and Pearson wanted to use the concept, he did not reject the meaning or utility behind it. In *Design of Experiments*, Fisher wrote the following:

By increasing the size of the experiment, we can render it more sensitive, meaning by this that it will allow of the detection of a lower degree of sensory discrimination, or, in other words, of a quantitatively smaller departure from the null hypothesis... We may say that the value of the experiment is increased whenever it permits the null hypothesis to be more readily disproved. (Fisher, 1966, pp. 21-22).

It is thus clear that Fisher placed great emphasis on sensitivity. He claimed the value of an experiment to be increased when its sensitivity is taken into consideration. Thus, it is clear that Fisher did not fundamentally disagree with the idea of Neyman and Pearson's power, but only that he did not approve of the concept being used in decision-making. As mentioned, it is likely that Fisher rejected the concept of power largely because he rejected the very foundations of Neyman and Pearson's Type II error, which forms the basis of power calculations. Fisher argued that calculations of power reflected the "mental confusion" (Fisher, 1955, p. 73) between technology and scientific inference. He did not, however, explain his reasoning for this. Surely, the concept of power as applied to statistical inference does not diminish the quality of scientific inference. I suspect that Fisher disliked the power advocated by Neyman and Pearson simply because they were applying it to quality control and to a decision-making process, and that because of this, it could not be used as part of scientific inference. That is, since Fisher was strongly opposed to Neyman and Pearson's decision-making for the purposes of scientific inference, it is likely that Fisher rejected the concept of power because it originated in this "decision-making" context. However, as has been shown by Cohen, power *does* have a place in scientific experiments, and in no way diminishes their scientific value. I argue also that Fisher rejected power *in totality* more because of the personal conflict with Neyman and Pearson than because of a genuine dislike for it. Surely, Fisher must have recognized the value of power in addressing his concerns of experimental sensitivity, not in the context of decision-making, but in the context of scientific inference. As Gigerenzer has said, "The concept of power makes explicit what Fisher referred to as 'sensitivity' " (Gigerenzer, 1993, p. 320). It is thus unlikely that Fisher would have whole-heartedly rejected Neyman and Pearson's power based solely on theoretical issues. Another possibility is that, since the development of power tables had to wait until the 1960s, Fisher may have been blind to the utility of such a tool in his statistical theory. He did however note other ways in which the sensitivity could be increased without necessarily increasing sample size. These included the structural organization of the experiment and refinements of technique. He mentioned these second to sample size increments.

Fisher's position with regards to sensitivity can be summarized to mean that each experimenter should take into consideration the sample size used when seeking a particular effect. Further, the experimenter should be aware of how

likely, even if only in a general sense, the null is to be rejected given its falsity. Fisher wanted us to consider the sensitivity of the experiment, something that many researchers still neglect in designing their research.

1.8. Fisher and the Hybridization of NHST

Now that an overview of the key components of Fisher's model of NHST has been given, it would do well to compare his model of significance testing to other "members" of the NHST hybrid, that of the Neyman-Pearson and Bayesian approaches to scientific inference. To fully understand the make-up of today's hypothesis testing as practiced by social scientists, and to learn why it can hardly be considered Fisherian, it is necessary to survey how the Neyman-Pearson and Bayesian camps interpret Fisher's seven components of the NHST model. Table 1 summarizes the primary components of the NHST model as advocated by Fisher and gives a contrast and comparison to that of the Neyman-Pearson paradigm and the Bayesian approach.

Early Fisher	Late Fisher	Neyman-Pearson	Bayesian	Current NHST
1. Forecast possible results *before* the experiment is executed, know beforehand the interpretation of various possibilities, should they arise	No change.	No comment	Provide a prior probability of the hypothesis under test, evaluate it in light of new data, and derive the posterior distribution of the estimated parameter.	Only partially satisfied through the specifying of usually only a single alternative hypothesis.
2 It is imperative that experimental units be assigned to conditions *randomly.* Only through randomization can error be properly estimated. Estimation of error presides over quantity of error.	No change.	Agree with Fisher· "Owing to the work of R.A. Fisher . One of the most important achievements of the English School is their method of planning field experiments known as the method of Randomized Blocks and Latin Square" (Neyman, J. et al , 1935, p. 109).	Agree with Fisher.	Usually not satisfied.
3. The sample is drawn from an *unknown hypothetical infinite population* We do not *know* the population from which our sample is drawn, we can only "imagine" the given population.	"The phrase 'repeated sampling from the same population' does not enable us to determine which population is to be used to define the probability level, for no one of them has objective reality" (Fisher, 1955, p 71)	The population being sampled is finite. We sample repeatedly from a given population " . a particular sample may be judged as likely to have been randomly drawn from a certain population, whose form may be either completely or only partially specified" (Neyman and Pearson, 1928, p 175).	The sample is considered observed, and hence fixed, which implies that an infinite hypothetical population makes little sense. (Personal communication with Jeff Gill, 2003)	N/A
4 There is only one hypothesis to be tested – the null hypothesis The null hypothesis *cannot* be established, yet it possibly can be "confirmed" or "strengthened."	No change	There are two competing hypotheses, the null and the alternative. A decision to accept one or the other must be made based on the outcome of the statistical test. "The Neyman Pearson position is that hypothesis testing demands a research hypothesis for which we can find support" (Cowles, 1989, p. 196).	The research hypothesis is assigned a probability conditional upon the observed data. There are no "accept/reject" decisions. "Multiple hypotheses can be simultaneously compared unlike frequentist likelihood ratio tests" (Gill, 2002, p 208) "A 'significant' result only means that the hypothesis of a null effect can be rejected, and a 'nonsignificant' result is nothing more than a statement of ignorance. On the contrary, Bayesian inference provides direct responses" (B. Lecoutre, 1998, p. 151)	Follows primarily the Neyman Pearson approach. Also incorporates quasi Bayesian notions of degrees of support for the alternative hypothesis based on low p-values in rejecting the null.

TABLE 1. — Comparison of Fisherian, Neyman-Pearson, Bayesian and the Current Approach to Hypothesis Testing

5. Using a significance level of 0 05 is convenient, but *not* mandatory.	$p = 0.05$ is a convenient significance level; the exact significance level should be reported in publication.	"In the long run of statistical experience the frequency of the first source of error [Type I error] . . can be controlled by choosing as a discriminating contour, one outside which the frequency of occurrence of samples . is very small – say, 5 in 100 or 5 in 1000" (Neyman and Pearson, 1928, p 177). " . if we take as the criterion of rejection p≤.01, let us say, we shall not accept samples for which Hypothesis A seems exceedingly improbable on any common sense grounds" (Neyman and Pearson, 1928, p. 184)	"Because the test is typically performed once on a set of social data in time and will not reoccur in the same fashion, the reported p value is not a long run frequentist probability" (Gill, 2002, p. 203) Rather than evaluating p(D/Ho), we should evaluate p(H/D), and assign it a probability, to be revised in light of further evidence.	A significance level of 0.05 or 0.01 is used, and contrary to Fisher, is usually deemed mandatory
6. Both statistically significant and non significant results should be published, as to yield a relative frequency from which a population could be shown to exist.	No change	No comment	No comment	Only "positive" results are typically published.
7. A researcher must consider the *sensitivity* of an experiment by either enlarging the number of repetitions (i e , sample size), or by qualitative methods	Power, or "Errors of the second kind" (Fisher, 1955) as advocated by Neyman and Pearson is inappropriate if the goal is scientific inference ". . . the value of the experiment is increased whenever it permits the null hypothesis to be more readily disproved" (Fisher, 1966, p. 22)	"It is not of course possible to determine [errors of the second kind] but making use of some clearly defined conception of probability we may determine a 'probable' or 'likely' form of it" (Neyman and Pearson, 1928, p. 177)	Bayesian methods are available for determining sample sizes Power calculations, as currently practiced, and being based on classical methods, are not needed (Rouanet, 1998, p. 63)	Historically, there has been a neglect for power (e.g , see Cohen, 1962; Rossi, 1990)

TABLE 1 (continued). — Comparison of Fisherian, Neyman-Pearson, Bayesian and the Current Approach to Hypothesis Testing

2. Contrasting Fisher's Model to Today's Hybridized Model

It is evident from Table 1 that today's social scientists practice anything but a pure Fisherian approach. Further, the model that is practiced, that is, the hybridized model, would likely be objectionable to each of the hybrid participants. Of course, the Fisherian and Neyman-Pearson approaches are more similar than that of the Bayesian approach. However, as noted by Poitevineau, it is indeed difficult to find "pure Neymaniens" (Poitevineau, 1998, p. 27) just as I argue it is an equal challenge to find pure Fisherians. What is more, as noted by Gill, even today's hybrid would likely not satisfy either camp: "Neither Fisher nor Neyman and Pearson would have been satisfied with the synthesis" (Gill, 1999, p. 653).

It is worthy now to discuss how today's hybridized model compares to the Fisherian model. As will be seen, components 1 through 3 are at least partially satisfied by today's model. However, components 4 through 7 constitute strong evidence for the claim that we cannot give Fisher's name to today's model of NHST. Using Table 1 as a guide, the following attempts to show how vastly different today's NHST model is from that of Fisher's proposed model, and further, vows to disentangle somewhat today's hypothesis testing hybrid. An empirical example is then given in which, as is typical of many published

research articles in the social sciences, the conflation of Fisherian, Neyman-Pearson and Bayesian approaches to NHST is unfortunately alive and well.

2.1. Forecasting of Results, Randomization, Hypothetical Populations

The first component, that of forecasting all possible outcomes and knowing beforehand the interpretation of each of these, can be said to at least be partially satisfied by today's NHST model. Most researchers design their experiments conscious of predicting possible outcomes, hence the process of specifying research hypotheses. However, whether today's investigators are prepared to interpret *all* possible results is open to debate. This may be largely due to the fact that there are often infinitely many possible outcomes when continuous measures are used. In relation to whether we follow Fisher's model, I score today's NHST a "yes-no" on this component; "yes" in that we specify our hypotheses beforehand in an effort to forecast possible results, and "no" in that we are often unprepared to account for results that deviate from our predicted outcomes. If results do not follow as expected, we are often left formulating post-hoc explanations to account for these unexpected findings. Fisher stressed that the theory must precede experimentation, and although today's researchers attempt to fulfill this requirement, it is many times left unfulfilled and we are ill-prepared to deal with unexpected findings.

Recall the second component of Fisher's model to be that of randomization. Subjects should be randomly assigned to treatment groups. Although the idea of random assignment is usually emphasized, in practice, it is seldom fulfilled. "Convenience samples" are often used and result in subjects not being randomly allocated to groups. A related problem is a lack of truly random samples and that of experimental generalizations that go far beyond the scope of the sample tested. Researchers often generalize their sample-based results to populations from which a random sample was *not* drawn. For instance, a sample of university students is only generalizable to a very narrow band of population parameters. However, we often see discussions generalizing to much wider populations. Further, the way in which subjects are recruited today would likely not satisfy Fisher. Even a most common and seemingly fair method of random selection, that of telephone number sampling, can only be generalized to the population consisting of those subjects who both have a phone, and are listed in the phone directory used in sampling. Because Fisher's methods were developed in the context of agricultural science, an analogy to this would be an investigator selecting those plots of land that are listed in the community's property listings. Such an investigation, while still useful, can only be generalized to those land estates listed. Perhaps land not noted was not listed because it was not fertile, and hence deemed not worthy of being listed. This is similar to the individual who is not recorded in the telephone directory because he suffers from major depression, which in turn results in him not being able to work, which in turn results in not having funds to afford a telephone. If the study were recruiting a sample to study the proportion of the population that suffers from major depression, the methodological problem is obvious. Even more methodologically unsound

is recruiting subjects by advertisement, as is often done in psychology, then attempting to generalize the study's results to a wider population than those subjects that served as volunteers. However, how often do we read results of the form, "These results suggest that male volunteers significantly differ from female volunteers on variable X (please note that these results can only be generalized to the 'volunteer-type subject)" ? As stated succinctly by Howell, "one person's sample might be another person's population" (Howell, 1989, p. 4). I argue that Fisher (1925, 1966) would have charged us with failing to recognize this. Fisher would advocate that although we are often randomly selecting, we are not randomly selecting from the population to which we generalize. However, Fisher may have well understood this to be a practical problem, and not a "true" departure from his original model.

2.2. We Randomly Select, Only From the Wrong Populations

The third component of Fisher's model consists of a hypothetical assumption that in a practical sense does not influence modern research customs to any significant degree. Fisher basically held that we cannot begin to specify the population from which our sample is drawn because we are unaware of it. If we were aware of it, then why would we have to sample in the first place? Hacking (1965) believes that the idea of hypothetical infinite populations contributes to unnecessary confusion. He argues that *chance set-ups* should be used to describe long-term frequency. He says: "However much they [*i.e.*, hypothetical infinite populations] have been a help,... hypothetical infinite populations only hinder full understanding of the very property von Mises and Fisher did so much to elucidate" (Hacking, 1965, p. 7). Later, Hacking continues:

One hopes our logic need not explicitly admit an hypothetical infinite population of tosses with this coin, of which my last ten tosses form a sample. Chance-set-ups at least seem a natural and general introduction to the study of frequency" (Hacking, 1965, p. 25).

By "chance-set-up", Hacking is referring to a system in which there are conducted experimental trials, of which each single trial is a member of a more complete class of possibilities. Thus for Hacking, having a class of possibilities is more enlightening and logical than having the population be infinite, as Fisher held. The population is a long-run frequency of *possibilities*, yet not infinite as Fisher would have. The debate of whether we sample from finite populations or hypothetical populations is a philosophical one, and shall be left to the philosophers of science to grapple with. The implication of either has no direct influence on how we practice hypothesis testing today. I would argue that few practicing researchers have given such a topic much thought, so although this component is a part of Fisherian NHST, whether it is even acknowledged by today's practitioners of NHST is unknown. Either way, in the world of significance testing difficulties, as the expression goes, "we have much larger fish to fry".

2.3. Hardly Fisherian

The following 4 components of Fisher's model are almost completely disregarded by today's researchers and journal editors. Hence, the following constitutes strong evidence that today's NHST is entirely dissimilar to the original Fisherian model, and shows how today's model represents a coarse hybridized blend of Fisherian, Neyman-Pearson, and Bayesian fundamentals.

2.3.1. Testing the Null vs. *Choosing Alternatives*

Component 4, that of positing only one hypothesis (the null) before an experiment, is not followed in the least by today's researchers, nor textbook writers. Today's procedure is that of setting up a null *and* an alternative hypothesis. Should the null hypothesis be rejected, the investigator decides on the *substantive* alternative as the most plausible argument, and even as evidential support for the obtained data. It is imperative to note that the substantive, or *conceptual* hypothesis, is the hypothesis that is held to best account for the data, given that the null is false. Usually, one substantive hypothesis is specified. The *statistical* alternative on the other hand, can be stated merely as "not the null", in that it suggests a distribution other than the null to account for the data. The primary difference between the statistical and the substantive hypotheses is that while the statistical hypothesis is simply a statement of "not the null", the substantive hypothesis constitutes an effort to explicitly account for the rejection of the null hypothesis. Given a rejection of the null, the statistical alternative is true. However, the substantive alternative may be only one of many hypotheses that is able to best account for why the null was rejected. In this respect, the substantive alternative serves as something of an "explanation" of why the null was rejected. As argued above, Fisher was skeptical in inferring an alternative hypothesis. The introduction of an alternative hypothesis is a Neyman-Pearson innovation and was applied in the context of decision-making – what Fisher would object to for the purposes of scientific inference. In Neyman-Pearson terms, the user of statistical methods needs to make a *decision* between two alternatives, not simply reject an unlikely hypothesis. As will be emphasized in the empirical example that concludes this article, understanding the decision-making logic of Neyman-Pearson hypothesis testing *versus* the scientific inference paradigm of Fisher's significance testing is *imperative* for a lucid interpretation of today's hybridized model. Although Fisher was not against using an alternative hypothesis when making decisions in industry, he was wholly suspicious of them for use in the field of pure scientific investigation. It is worth quoting Fisher (1966) extensively here for an acute sense of his position on what he called "Acceptance Procedures":

The situation is entirely different in the field of Acceptance Procedures, in which irreversible action may have to be taken, and in which, whichever decision is arrived at, it is quite immaterial whether it is arrived at on strong evidence or on weak. All that is needed is a Rule of Action which is to be taken automatically, and without thought devoted to the individual decision. The procedure as a whole is arrived at by minimizing the losses due to

wrong decisions, or to unnecessary testing, and to frame such a procedure successfully the cost of such faulty decisions must be assessed in advance; equally, also, prior knowledge is required of the expected distribution of the material in supply. In the field of pure research no assessment of the cost of wrong conclusions, or of delay in arriving at more correct conclusions can conceivably be more than a pretence, and in any case such an assessment would be inadmissible and irrelevant in judging the state of the scientific evidence; moreover, accurately assessable prior information is ordinarily known to be lacking. Such differences between the original situations should be borne in mind whenever we see tests of significance spoken of as "Rules of Action". *A good deal of confusion has certainly been caused by the attempt to formalise the exposition of tests of significance in a logical framework different from that which they were in fact first developed* [emphasis added]. (Fisher, 1966, pp. 25-26).

In the above passage, Fisher left little doubt of how he feels with regard to his tests being used in the field of so-called "Acceptance Procedures" or used as "Rules of Action". From his earliest inception of significance testing to his death, Fisher adamantly held that these procedures were *not* to be used for the purpose of judging scientific evidence. Today however, researchers continue to employ these procedures as a model for establishing scientific evidence. I would venture to suggest that on this basis alone, Fisher would want little to do with today's NHST.

2.3.2. Sacred Significance

Component 5, that of the "convenient" use of a significance level of 0.05, is totally dismissed by most researchers and journal editors. Fisher stated that the 0.05 level of significance is "convenient", not gospel. Today, we seldom find an experimental study in which the 0.05 level has not been used. An exception to this occurs when we read a significant result at the 0.01 level. Rarely, if ever, do we find a significance level of over 0.05, even if the difference is slight. Journal editors have been found to be quite rigid in their demand for the null to be rejected at least at 0.05 (*e.g.*, see Melton, 1962). As noted previously, Fisher, later in his career, required experimenters to state precisely the significance level when reporting results, the *exact* probability. Huberty (1993) found that many recent textbook authors suggested choosing a significance level *prior* to data collection. The data may yield a probability of 0.03, however many of today's journals list it as merely lower than 0.05, despite the fact that the most recent edition of the American Psychological Association Publication Manual (2002) cites both ways as acceptable. Therefore, the combination of adhering to a rigid probability level (counter early Fisher), and not reporting exact significance levels (counter late Fisher), are two important elements in significance testing that run opposite to Fisher's recommendations. Furthermore, as noted by Huberty, despite Fisher's rejection of a fixed level of significance for all experiments (see Fisher, 1959), some researchers still cite him as support for their choice of the 0.05 level of significance. The issue surrounding significance levels is perhaps the most compelling reason why we can hardly attribute

today's model to Fisher – we simply do not do significance testing as Fisher prescribed.

2.3.3. Publishing Positives Only

Component 6, that of publishing both significant *and* non-significant results, is also largely ignored by today's researchers and journal editors. As noted by Gigerenzer *et al.* (1989), Fisher would have wanted for us to be publishing both significant and non-significant results so that a relative comparison between the two groups could be made as to ascertain the existence of a given phenomenon. Indeed, only by a comparison of the two groups can we claim with any kind of authority that a phenomenon exists. Today's methods overlook the importance of accounting for negative results, and these are typically not published. There have been some efforts to change this state of affairs however. Neher (1967) for instance, argued that psychological research falls victim to what is referred to as "probability pyramiding", a process by which the significant outcomes are reported more faithfully than the insignificant outcomes. As Neher demonstrated:

In an extreme case, for example, 20 different analyses might be done; 19 may be insignificant at the 5 per cent level and the 1 analysis that is significant may be reported as a finding. Obviously, this is about what one would expect to find on the basis of chance alone, so that this is likely to be a spurious finding (a Type I error). (Neher, 1967, p. 257).

This, of course, is a major consequence of not counting negative results. A way around this problem would be to account for both significant and non-significant results, thus being able to directly compare the frequency of each. Smart (1964) has also noted the importance of negative results in research. He found that only 9 per cent of the aggregate of papers published were projecting negative findings. I would argue that today, this figure is probably even lower. Journal editors do not want to publish research that barely fails to meet the 0.05 probability level, never mind being outright negative! Fisher would disagree with today's "positive-only" publishing. Also, implicit in point 6, is the idea that experiments should be replicated. Only by replicating an exact experiment could we possibly arrive at a ratio of "positive vs. negative" results. Today, researchers do neither. Negative results are not accounted for, and *exact* replication is almost non-existent. Fisher would not approve.

2.3.4. Sensitivity and Sample Size

Component 7, that of the sensitivity of an experiment, is again, largely disregarded by today's community of researchers. Fisher wanted us to consider the sensitivity of an experiment in being able to reject the null hypothesis. As mentioned earlier, Fisher advocated a conceptually similar construct to the Neyman-Pearson idea of power. Today, researchers rarely concern themselves with the sensitivity or the power of their research. Power in most textbooks is seldom discussed adequately, if at all. Furthermore, as with many other statistical concepts, power is poorly defined in some texts (Brewer, 1985).

Again, it is tempting to argue that Fisher's sensitivity is similar to Neyman and Pearson's power. Except for the idea of estimating the Type II error

rate, Fisher's idea of sensitivity is very similar to the *idea* of power. Perhaps Fisher would have been more accepting of power after Cohen devised ways of actually calculating it. Regardless, it is a well-known fact that today's researchers devote little attention to either the sensitivity or the power of their experiments. Power surveys have suggested the calculation of power to be almost non-existent in journals, and when calculated, to be quite low (Cohen, 1962; Rossi, 1990). Fisher wanted us to think of sensitivity as a way of improving our experiments, as he said "it [sample size increase] will allow... of a quantitatively smaller departure from the null hypothesis" (Fisher, 1966, p. 22). Researchers today seem almost indifferent as to the number of subjects they use. Instead, they use rough guidelines and hope for the best. This has resulted in some post-hoc sample size studies to show an almost impossibility of rejecting the null hypothesis in many studies (Tversky Kahneman, 1971). Whether it is Fisher's sensitivity, or Neyman and Pearson's power, either is seldom addressed in today's research, hence again further distancing today's model from Fisher's original model of significance testing.

3. Fisherian Felony ?

Because many still believe we employ a Fisherian model in social statistics, this has resulted in unjust criticisms towards Fisher – laying blame with Fisher for things he never once defended nor supported. It is appropriate to cite a typical instance where Fisher was wrongly faulted for ideas that were not his own. The above overview of NHST will help make obvious the misattribution in the following accusation.

Meehl is perhaps the harshest with Fisher. He blames Fisher outright for our misuse of significance testing. He argues:

I suggest to you that Sir Ronald has befuddled us, mesmerized us, and led us down the primrose path. I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas [of social science] 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 correct, the over-reliance on refuting the null *is* detrimental to our discipline – but this is not Fisher's fault. First, Fisher did not *lead* us anywhere, we lead ourselves astray. Meehl's critique is that much less credible when one considers that Fisher *never* recommended his procedures for social science. I dare ask how he could have led us down the primrose path when he never even suggested we follow him! Meehl automatically (and mistakenly) associates modern significance testing with Fisher's name, and if I've shown nothing else, doing this is misleading, and hardly fair to Fisher's legacy.

4. Hybrid Hypothesis Testing: An Empirical Example

It would do well in concluding this article to provide an empirical example of how Fisherian NHST is incorrectly conflated with both Neyman-Pearson and Bayesian logic, in what has become a hybrid of hypothesis testing procedures. A typical instance of the hybrid occurs when a researcher rejects a null hypothesis (Fisher), accepts and concludes an alternative hypothesis (Neyman and Pearson), and implicitly treats smaller p-values as increasingly stronger support for the alternative hypothesis (quasi-Bayesian). The present example comes from the literature on marital satisfaction. In a study by Snow and Compton (1996), researchers were interested in the relationship between religious affiliation and such things as marital adjustment and marital communication. A first research hypothesis was that membership in a fundamentalist church would be related to marital satisfaction, marital communication, or both. In testing the null, that of no relation between membership and marital satisfaction or communication, it is clear the authors believe that a non-significant result supports that of no relationship. That is, a non-significant result stands to *support* the null hypothesis. With regard to this first research hypothesis, they conclude:

One-way analyses of variance between groups were nonsignificant for means on the Dyadic Adjustment Scale, the Marital Communication Inventory, and all of their subscales and, therefore, *did not support the hypotheses* [emphasis added] that membership in a fundamentalist church is a significant factor in either marital satisfaction or communication (Snow and Compton, 1996, p. 982).

There is an unequivocal conflation of Fisherian, Neyman-Pearson and arguably Bayesian hypothesis testing present in the researchers' interpretation of their statistical test. The authors claim that a nonsignificant difference fails to support the research hypothesis. However, according to Fisherian significance testing, a nonsignificant result establishes no such thing. Fisher would say that failing to reject the null simply means there is insufficient evidence against it (*i.e.*, the null, not the alternative), and that the experiment has failed to produce a significant result – end of story. The above interpretation is more in line with Neyman-Pearson logic, in which the researcher must seemingly "choose" between two alternatives. However, even a quasi-Bayesian interpretation can be drawn from the researchers' misuse of significance testing. If we ignore for a moment the "cliff effect",[2] then it would appear that had the researchers obtained significance, the strength of the misattributed support for the alternative hypothesis might be implied by the smallness of the p-value. That is, by incorrectly crediting strength to the alternative given a rejection of the null by an arbitrary p-value, it is likely the researchers' *belief* in the alternative would increase proportional to the smallness of the p-value. Indeed, as noted by Lecoutre, "the significance test is one of the elements

2. The so-called "cliff effect" describes the tendency of researchers to dramatically lower their confidence in research findings as the probability of the obtained statistic rises above 0.05 (Rosenthal and Gaito, 1963). However, Poitevineau and Lecoutre (2001) have found the effect to be somewhat exaggerated.

of a whole rhetoric for the presentation of results which is currently used by researchers to strengthen their arguments and convince their colleagues of the value of their results" (Lecoutre, 1998, p. 78). The authors, in their making explicit the claim of testing the alternative, and thereby mistakenly ascribing strong evidence against the null as strong support for the alternative, also imply that evidential support for the alternative can be measured by the smallness of the p-value. Such an interpretation is an example of what Gigerenzer has called the "Bayesian Id" (Gigerenzer, 1993, p. 330). Indeed, as Cohen (1994) has remarked, most users of NHST evaluate hypotheses wanting to be Bayesian about them. However, due to numerous social, political and editorial influences, Bayesian statistics, even though regarded as superior to NHST by most methodologists, have been slow to progress in the social sciences: "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" (Lecoutre, Lecoutre and Poitevineau, 2001, p. 401). Unfortunately, what most often occurs, as evident in the present example, is a serious conflation of Fisherian, Neyman and Pearson, and Bayesian hypothesis testing.

In further discussing the results of a multiple regression, the authors also mistakenly claim evidence for the null hypothesis, when no such conclusion is warranted: "These results suggested that membership in a fundamentalist Protestant church was not a predictor of either marital satisfaction or satisfaction with marital communication" (Snow and Compton, 1996, p. 982). Again, by the absence of a significant result, the authors seem content in substantiating the null hypothesis, that of no relationship between membership and marital satisfaction. A pure Fisherian interpretation would be simply that we have no statistical evidence to suggest there to be a relationship between membership and marital satisfaction. However, nowhere would this preclude there being one, only that *our experiment has failed to find one.* What is more, it is doubtful that the conclusion reached by the authors lends itself even to a pure Neyman-Pearson interpretation either since it is taken out of the context of the decision realm in which Neyman-Pearson hypothesis testing arose. That is, the authors seem to *conclude* that there is no relationship, apparently as a scientifically credible finding. However, as is well known in the Neyman-Pearson hypothesis testing logic, "accepting" the alternative is made as part of a *decision* against a competing null. That is, if one *had* to choose one option over the other, in given time and space, such as would be the case in quality control testing for instance, then given a failure to reject the null hypothesis, one would choose the alternative. *But, this is a decision, not a scientific claim of no relation.* Distinguishing between these two ideas is of fundamental importance in comparing Fisherian to Neyman-Pearson hypothesis testing, and must be understood for an acute appreciation of the modern hypothesis testing hybrid.

5. Conclusion

In closing, I have argued I hope successfully, that to ascribe Fisher's name with today's NHST is nothing short of an academic misdemeanor. Today, researchers do something different from what Fisher once proposed, and to call today's statistical practices "Fisherian" does not do justice to the statistical genius. Today's model is much too hybridized, misused, and misunderstood to be attributed with the likes of a statistical pioneer as Fisher. Therefore, with Cowles (1989), I hope you will concur – to call today's procedures "Fisherian" is indeed likely to cause Sir Ronald to unduly shuffle in his tomb. Allow the Master to rest, once and for all.

References

AMERICAN PSYCHOLOGICAL ASSOCIATION. (2002). *Publication manual of the American Psychological Association (5th ed.)*. Washington, DC: Author.

BAKAN D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 423–437.

BORING E. G. (1919). Mathematical vs. scientific significance. *Psychological Bulletin*, 16, 335–338.

BREWER J. K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions ? *Journal of Educational Statistics*, 10, 252–268.

CARVER R. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, 61, 287–292.

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

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

COHEN J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997–1003.

COWLES M. (1989). *Statistics in psychology: an historical perspective*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

FISHER R. A. (1990/1925). *Statistical methods for research workers*. London: Oliver and Boyd (14th edition 1973 reprinted, Oxford University Press, 1990).

FISHER R. A. (1935a). *The design of experiments*. New York: Hafner Publishing Company.

FISHER R. A. (1935b). The logic of inductive inference. *Journal of the Royal Statistical Society*, 98, 39–82.

FISHER R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Society*, 17, 69–78.

FISHER R. A. (1956). *Statistical methods and scientific inference*. Edingburgh: Oliver & Boyd.

FISHER R. A. (1959). *Statistical methods and scientific research (2nd ed.)*. New York: Hafner.

FISHER R. A. (1966). *The design of experiments (8th ed.)*. New York: Hafner Publishing Company.

FISHER BOX J. (1978). *R. A. Fisher: the life of a scientist.* New York: John Wiley & Sons.

GIGERENZER G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (eds.), *A handbook for data analysis in the behavioral sciences: methodological issues (pp. 311-39).* Hillsdale, NJ: Lawrence Erlbaum Associates.

GIGERENZER G., SWIJTINK Z., PORTER T., DASTON L., BEATTY J., KRÜGER L. (1989). *The empire of chance: how probability changed science and everyday life.* New York: Cambridge University Press.

GILL J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly*, 52, 647–674.

GILL J. (2002). *Bayesian methods: a social and behavioral sciences approach.* London: Chapman Hall/CRC.

GOSSET W. S. (1937). "Student". Comparison between random and balanced arrangements of field plots. *Biometrika*, 29, 363–79.

HACKING I. (1965). *Logic of statistical inference.* London: Cambridge University Press.

HOWELL D. C. (1989). *Fundamental statistics for the behavioral sciences (2^{nd} edition).* Boston: PWS-Kent Publishing Company.

HUBERTY C. (1993). Historical origins of statistical testing practices: the treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education*, 61, 317–333.

KENDALL M. G. (1943). *The advanced theory of statistics (Vol. 1).* New York: Lippincott.

LECOUTRE B. (1998). From significance test to fiducial Bayesian inference. In Rouanet, H., Bernard, J. M., Bert, M. C., Lecoutre, B., Lecoutre, M. P., Le Roux, B. (1998). *New ways in statistical methodology: From significance tests to Bayesian inference.* Berne: Peter Lang.

LECOUTRE M. P. (1998). And... what about the researcher's point of view ? In Rouanet, H., Bernard, J. M., Bert, M. C., Lecoutre, B., Lecoutre, M. P., Le Roux, B. (1998). *New ways in statistical methodology: From significance tests to Bayesian inference.* Berne: Peter Lang.

LECOUTRE B., LECOUTRE M. P., POITEVINEAU J. (2001). Uses, abuses and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable ? *International Statistical Review*, 69, 399–417.

MEEHL P. (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, 553-7.

NEHER A. (1967). Probability pyramiding, research error and the need for independent replication. *The Psychological Record*, 17, 257–262.

NEYMAN J., PEARSON E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference (part 1). *Biometrika, 20A*, 175–240.

NEYMAN J., IWASZKIEWICZ K., KOLODZIEJCZYK St. (1935). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, 2, 107–180.

POITEVINEAU J. (1998). Méthodologie d'analyse des données expérimentales... *Études de la pratique des tests statistiques chez les chercheurs en psychologie,*

approches normative, prescriptive et descriptive. Thèse Univ. De Rouen. Sous la dir de B. Lecoutre.

POITEVINEAU J., LECOUTRE B. (2001). Interpretation of significance levels by psychological researchers: The .05 cliff effect may be overstated. *Psychonomic Bulletin & Review*, 8, 847–850.

ROUANET H., BERNARD J. M., BERT M. C., LECOUTRE B., LECOUTRE M. P., LE ROUX B. (1998). *New ways in statistical methodology: From significance tests to Bayesian inference.* Berne: Peter Lang.

ROSENTHAL R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86, 638–641.

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

ROSSI J. S. (1990). Statistical power of psychological research: What have we gained in 20 years ? *Journal of Consulting and Clinical Psychology*, 58, 646–656.

SMART R. G. (1964). The importance of negative results in psychological research. *The Canadian Psychologist*, 5, 225–232.

SNOW T. S., COMPTON W. C. (1996). Marital satisfaction and communication in fundamentalist protestant marriages. *Psychological Reports*, 78, 979–985.

TVERSKY A., KAHNEMAN D. (1971). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.

VENN J. (1876). *The logic of chance.* London: Macmillan and Co.