M@n@gement

ISSN: 1286-4892

■ Ababacar MBENGUE 2010

Should we burn the statistical significance tests?
M@n@gement, 13(2), 99 - 127

accepted by Louis Hébert

M@n@gement est la revue officielle de l'AIMS

AIMS Association Internationale de Management Stratégique

M@n@gement is the official journal of AIMS

M@n@gement vol. 13 no. 2, 2010, 99-127

Should we burn the statistical significance tests?

Ababacar Mbengué

Reims Management School
ababacar.mbengue@univ-reims.fr

Inference plays a central role in management research as researchers are frequently led to draw conclusions or make generalizations from their observations or results. In many cases, they are able to do this rigorously through inferential statistics, which is the process of inference whereby the statistician tests the generalization of information collected in a sample to the entire population the sample is from. Statistical tests are thus at the heart of inferential statistics and, consequently, the process of inference. However, since they were first developed, statistical significance tests have been the object of sharp and repeated criticism regarding both their nature and their role (Nickerson, 2000). Such criticism has been longstanding in virtually all disciplines, with the notable exception of management that is just beginning to address the issue (Mbengue, 2007, Schwab & Starbuck, 2009). The main purpose of this paper is to provide researchers in management with clear information about the controversy surrounding statistical significance tests, to detail the content and issues and, most importantly, to offer recommendations for improving the testing of hypotheses and beyond, in other words, the process of statistical inference in management research.

Keywords: *methodology, inferential statistics, statistical tests, statistical significance*

INTRODUCTION

Since their inception, statistical significance tests, better known as Null Hypothesis Significance Tests (NHST), have been subject to numerous criticisms concerning both their nature and role. And while some authors have provided a defence of these tests (Hagen, 1997; Mulaik, Raju, & Harshman, 1997; Wainer, 1999), many others have called for their outright abolition (Hunter, 1997; Gill, 1999; Armstrong, 2007a, 2007b). In fact, few statistical methods have been as strongly and persistently criticized as NHST (Morgan 2003), which Rozeboom (1960: 416) denounced as a "fallacy", Bakan (1966: 436) called an exercise in "mindlessness in the conduct of research," Carver (1978: 397) a "corrupt form of the scientific method,» Krueger (2001: 16) a "flawed method," while Schmid & Hunter (2002: 66) described them as «disastrous.» Yet, despite these numerous criticisms, NHST continue to be as popular as misused (Finch, Cumming, & Thomason, 2001; Morgan, 2003).

In fact, numerous studies from disciplines as diverse as psychology, sociology, marketing, accounting, education science, political science, ecology, forecasting, psychiatry, etc. consistently show that researchers ignore the most basic aspects of NHST (Nelson, Rosenthal, & Rosnow, 1986; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993; Mittag & Thompson, 2000). So, should we burn the NHST?

To answer this question, we will briefly describe the general logic of NHST, present the main criticisms and then make recommendations to improve the testing of hypotheses and beyond, in other words, the process of statistical inference in management research.

GENERAL LOGIC OF STATISTICAL SIGNIFICANCE TESTS

STATISTICAL INFERENCE

The process of inference plays a central role in researchers' work. The latter are frequently called upon to interpret results, draw conclusions or make generalizations from their observations. For example, to what extent can results obtained from a study involving a limited number of cases, individuals, social groups, organizations, processes, etc. also be applied to other cases, individuals, social groups, organizations and processes not included in the initial research? This is an essential feature of the scientific process because the interest of a specific piece of research is often highly dependent on the more or less generalizability of the findings, although of course, no research is expected to come up with universally applicable results. It is, in fact, quite rare for researchers to merely describe situations, phenomena or raw data, but they frequently make inferences to generalize their results. There are generally two main forms of inference: 1) theoretical inference or analytical generalization (Kennedy, 1979; Yin, 1984; Firestone, 1993; Smaling, 2003), which aims to generalize theoretical propositions based on logi-

cal reasoning, and 2) statistical inference, which is a form of generalization based on the properties of mathematical statistics. While each of these two forms of inference occupies an important place in scientific research, statistical inference is at the heart of NHST in that it allows the researcher to test the generalization of information collected in a sample to the entire population that the sample is from. This statistical form of inference involves two classes of tools: 1) the statistical significance tests which will be discussed in this article and 2) the estimation methods whose presentation is beyond the scope of this paper.

In any event, any statistical generalization of research findings only has meaning in probabilistic terms. That is, statistical inference does not lead to definite judgments, but rather, to more or less likely judgments, which in turn means that it cannot completely eliminate the risk of error associated with the choice of giving validity to a phenomenon that goes beyond the narrow context in which it has been observed. However, statistical inference is of very real interest in that it can control the risk of error. More exactly, it manages to make a theoretically accurate estimation, indicating the likelihood that a researcher would be wrong in generalizing his or her study findings. To this end, statistical inference uses descriptive statistics and probability theory in order to study the laws and regularities governing random phenomena. This theory enables statistical inference to determine the likelihood that a phenomenon observed in a sample is only due to random sampling, even if it is absent from the entire sample population.

Statistical inference is made using statistical hypothesis testing. A statistical hypothesis is a quantitative statement about the parameters of a population (Baillargeon & Rainville, 1978). A population parameter is a quantitative aspect of this population such as the mean, variance, percentage or any specified quantity regarding the said population. The parameters of a population are generally unknown. However, it is possible to statistically estimate them with a sample from the population. To make a statistical inference, researchers must translate their research hypothesis into a statistical hypothesis. For example, a research hypothesis may be that for a given population of firms, i.e. those which adopted Total Quality Management are performing better than those which did not. Since moving from a research hypothesis to its statistical test necessitates the translation of the research hypothesis into a statistical hypothesis, the details of such a translation are consequently crucial for the validity of an inference on the research hypothesis. However, statistical test theories only deal with the statistical hypothesis, and the question of the relevance of the translation of the research hypothesis via a statistical hypothesis is not their responsibility (Poitevineau, 1998).

A statistical hypothesis is traditionally presented in the dual form of a first hypothesis called a null hypothesis and a second one called an alternative hypothesis. The null hypothesis is generally the situation of no change or deviation from the status quo, or the absence of difference between parameters, hence the term ‘null hypothesis’ (Kanji, 1993 Dodge, 1993). Very often, the research objective is to refute the null

hypothesis in favor of the alternative hypothesis (Dodge, 1993; Sincich, 1996; Zikmund, 1994). The alternative hypothesis is the one the researcher wishes to establish, what he or she believes in (Sincich, 1996). The null hypothesis and the alternative hypothesis are incompatible and describe two complementary states of nature. The null hypothesis is usually denoted by H0 and the alternative hypothesis by H1 or Ha. It should be noted that the statistical tests are designed for the refutation and not the confirmation of hypotheses. In other words, these tests have neither the ambition nor the power to prove hypotheses and can only show that a hypothesis is unacceptable because the associated probability level is too low (Kanji, 1993).

Although they are generally presented as a unified theory, NHST are actually a hybridization of two competing paradigms: Fisher's significance test and Neyman and Pearson's hypothesis test (Hubbard & Bayarri 2003). The distinction between these competing schools of thought does not seem to have occupied a major place in the use of NHST debate, although some authors explain a good deal of the limitations of NHST by the original hybridization (Harlow, 1997; Gill, 1999; Hubbard & Bayarri 2003).

Tests and statistical errors

Assessing a statistical hypothesis' validity is done by using a statistical test performed on data from a representative sample of the population studied. This statistical test is a procedure that leads to the rejection, or failure to reject, of a hypothesis, generally the null hypothesis. The specific form of the statistical tests depends on the number of populations involved (one, two or more). For a statistical test on a single population, it is important to know whether the value of a population parameter θ is identical to an expected value. In this case, the null hypothesis is an assumption about the expected value of this parameter and is usually expressed as follows:

$$H0: \theta = \theta 0,$$

where θ is the population parameter and $\theta 0$ the assumed value of the unknown parameter θ to be estimated.

The alternative hypothesis, on the other hand, posits the existence of a difference or inequality. For example, we can assume superior performance for firms that plan formally. In such cases, the statistical test to be performed is called a one-tailed or unidirectional test. If the hypothesis is simply that of a performance difference without more precision, there should be a two-tailed or bidirectional test. It thus appears that the alternative hypothesis can take different forms:

- $H1: \theta > \theta 0$ (one-tailed or unidirectional)
- $H1: \theta < \theta 0$ (one-tailed or unidirectional)
- $H1: \theta \neq \theta 0$ (two-tailed or bidirectional)

Statistical tests are performed in order to make a decision, namely to reject or not the null hypothesis, H0. But because the decision is based on partial information following observations on a sample population only, there is obviously potential for error (Baillargeon & Rainville, 1978; Sincich, 1996; Zikmund, 1994).

There are two types of errors in statistical tests: type I error, also known

as an «error of the first kind», whose probability is denoted α., and type II error, also known as an «error of the second kind», whose probability is denoted β. Sample observations can lead to a rejection of the null hypothesis H0, even though the population meets the conditions for this hypothesis. Conversely, the null hypothesis H0 may not be rejected following the sample observations even though the population qualifies for the alternative hypothesis H1. A Type I error can occur only in cases where the null hypothesis is rejected. Similarly, a type II error can occur only in cases where the null hypothesis is not rejected. Therefore, either researchers are not guilty of an error or they are guilty, but only of one type of error. They cannot commit both types of error simultaneously. The power of a statistical test is the probability that the test will reject a false null hypothesis (i.e. that it will not make a Type II error). As power increases, the chances of a Type II error decrease. The probability of a Type II error is referred to as the false negative rate (β). Therefore power is equal to 1 – β. Most statistical software provides the probability associated with the observed value of the statistic calculated or the observed level of significance, better known as the p-value. This is the probability, calculated under the null hypothesis, of obtaining a result as extreme as the value obtained by the researcher from the sample (Dodge, 1993). The null hypothesis H0 will be rejected if the p-value is below the significance level α (Sincich, 1996).

Specifically, NHST provides an assessment of the risk of error to which researchers are exposed when generalizing their research findings. For example, researchers who find a correlation between size and knowledge management practices in a sample of sixty French companies may wish to determine if they can reasonably conclude that a similar phenomenon is also present in the entire population of French firms. To this end, they conduct a statistical test to ascertain to what extent (with which risk of error) the correlation found in the sample of sixty firms can be generalized to the entire population of French firms. Specifically, NHST allow researchers to determine the probability of whether a result equal to or greater than that which has been obtained on the sample is due solely to random sampling, and to assess the risk of error in making a sweeping generalization of the result of the sample to the whole population.

CRITICISMS AGAINST STATISTICAL SIGNIFICANCE TESTS

Basically, NHST pose three types of problems: 1) statistically, they have inherent weaknesses that are too often ignored by researchers, 2) technically, the dominant use of these tools is generally inadequate, and 3) philosophically, they raise the more general issue of the limits of inference and are ultimately based on the unproven belief that understanding the past can predict the future (Cox, 1958; Morgan, 2003).

STATISTICAL PROBLEMS

About the null hypothesis

Several prominent statisticians have long recognized that the null hypothesis of no difference is never true in the population as a whole (Nunnally, 1960; Tukey, 1991). Among the many examples of null hypotheses, we could mention: a frequency or proportion of zero, a zero correlation, a regression coefficient equal to zero, a tie between two or more averages, etc. According to Schwab & Starbuck (2009), such examples are very common in management research. Many critics of NHST call into question the very usefulness of a point null hypothesis (H0), in other words, a hypothesis attributing a specific value to the parameter and not an interval. Indeed, such an assumption is almost always false (Oakes, 1986; Tukey, 1991; Morgan, 2003). Thus, any difference, even infinitesimal, becomes statistically significant provided that the sample size is large enough. For example, for a mean difference test between two groups, any not strictly zero difference may be 'made' significant as long as the group size is large enough. Let's consider the example of a research experimentation testing the hypothesis (actually true) that a specific training program increases employees' productivity. The experimental group (employees who have been trained) and the control group (employees who have not been trained) then represent two different populations. The experimental group presents an average performance of 76% and a standard deviation of 4.3 and the control group an average performance of 72% with a standard deviation of 4.3. It can be shown that for representative samples of both populations, a size of at least 11 employees per group is needed to detect a statistically significant difference in performance between the two groups. Thus, if the study involves fewer than 10 employees per group, it will yield a statistically non significant result, whereas if the number of employees is over 10 it will yield a statistically significant result. In this example, the researcher can directly determine the statistical significance of the results only by controlling the size of the sample. Another way of looking at things is to start from the population rather than the samples. It is very likely that, at population level, the null hypothesis is false (there will be a difference, even infinitesimal). Therefore, any test will lead to a significant result and the information provided by the test is virtually zero. This has led many critics of NHST to conclude that collecting data and carrying out such tests whose results are known in advance is futile (Meehl, 1990; Krueger, 2001; Morgan, 2003).

An anti-scientific approach?

The emphasis put on the null hypothesis by NHST (chance, no difference, status quo, lack of effect, etc.) tends to weaken the scientific process. Indeed, a scientific approach essentially consists of comparing data against the research hypothesis (H1), and when the data appear inconsistent against H1, to consider alternative hypotheses (including that of chance). However, in the case of NHST, the hypothesis of chance (H0) is usually put forward and tested first, regardless of its (un) scientific interest (Carver, 1978, 1993; Cohen, 1994). Therefore, the research hypothesis will not even be considered if the test is not significant, even if it is highly compatible with the data. If the choice to assimilate the null hypothesis to random situations, inexistence of difference and effect, etc. has many technical advantages such as simplification of certain statistical calculations, the fact remains that it distracts from the main concern of the researcher which is the research question (usually H1). However, when NHST are used, it is as if only two models existed or deserved to be considered: the null hypothesis and the alternative hypothesis (often the research hypothesis). This is unrealistic since several other alternative models are possible (Rozeboom, 1960; Lindsay, 1995; Morgan, 2003). Let's take the example of a research hypothesis that «the distribution of stock options to executives increases their loyalty.» The null hypothesis is then expressed as follows: «the distribution of stock options to executives does not change their loyalty.» A test for difference of means between the two samples of executives who received and did not receive stock options then reveals a much higher retention rate for executives who received stock options, resulting in a statistically significant difference in the mean. The rejection of the null hypothesis here is worth accepting, or at least «not rejecting» the research hypothesis that the increased retention is due to the awarding of stock options. However, several other phenomena may be the cause of the observed mean difference in the awarding of stock options (the research hypothesis), such as methodological bias or measurement error, or variables such as type of business, seniority in the company or the executives' personal profile. As Carver noted (1978), when rejecting the null hypothesis, researchers should be able to reject all rival explanations to the research hypothesis that might also explain its rejection. In other words, rejecting the null hypothesis means admitting that «there is not nothing going on» but does not indicate that what is happening is exactly what the research hypothesis states.

Null hypothesis, alternative hypothesis and research hypothesis

When the research hypothesis (which the researcher is really interested in) leads to an accurate prediction of the parameter, it is possible to identify it with the null hypothesis rather than the alternative hypothesis. This is particularly true with model validation. An illustration can be found in causal model testing, such as a causal model explaining executive loyalty by the granting of stock options. Another example of a research hypothesis corresponding to the null hypothesis is that of the hypothesis that organizational survival is a random phenomenon. This

hypothesis is based on the results of empirical work conducted in the analytical context of population ecology that showed no difference in the survival rates of either young and old organizations or small and large organizations (Schwab & Starbuck, 2009). But whatever the choice adopted by researchers for the research hypothesis (H0 or H1), difficulties remain: if we identify the research hypothesis with the alternative hypothesis (classic case), then, as already mentioned, we just need to choose a large enough sample to be sure of obtaining a favorable outcome (significant). Conversely, if we identify the research hypothesis with the null hypothesis, we are faced with the following dilemma (Cohen, 1992; Poitevineau, 1998, Martinez-Pons, 1999):

- to build a highly sensitive research design (for example by choosing a large sample) which leads to the previous case of almost certain rejection of the hypothesis when it may be a very good approximation of the reality, and even the best available one;
- to build a research design with poor sensitivity (e.g. by selecting a small sample) which allows easy and artificial corroboration of the research hypothesis.

Probability of the hypothesis or probability of the data?

NHST provide a probability of observed data that is conditional on the truth of the null hypothesis, or Pr (Data|H0). Yet what really interests the researcher is in fact the probability of the hypotheses (H0 and/or H1), conditional on observed data Pr (H0|Data) or Pr (H1|Data). This unnatural and unintuitive position consequently leads to frequent misinterpretations as mentioned by several authors (Carver, 1978; Cohen, 1994; Gill, 1999). Many researchers interpret the observed results Pr (H0|Data) as the conditional probability of the null hypothesis, namely Pr (H0|Data). Yet the values of these two probabilities can be very different from one another, i.e., 0.005 to 0.82 as in one famous case described by Falk & Greenbaum (1995). Let's consider the example of a study that aims to test the hypothesis that the existence of a strategic planning department in SMEs improves their performance. The mistake is to confuse the two probabilities: 1) Pr (superior performance|presence of a strategic planning department) is the probability of observing superior performance in the presence of a strategic planning department in the SMEs and 2) Pr (presence of a strategic planning department|superior performance) which is the probability of the presence of a strategic planning department in SMEs with high performance. These two probabilities may of course be very different.

α is arbitrary, p-value is ambiguous

The selected significance level α (usually 5% but sometimes 10% or 1%, etc.) is totally arbitrary, yet it leads directly to determining the rejection or non-rejection of hypotheses. Let's take the example of comparing a mean of one sample to a given value. We have a sample of 144 observations. The average found in this sample is m = 493. The estimated standard deviation on the sample is s = 46.89 and the observed level of significance or p-value is p = 0.07. In this example, we will

not reject the (null) hypothesis that the population mean is μ0 = 500 if the selected significance level α is 5%, whereas we would reject it if the chosen significance level α is 10%, which is clearly arbitrary.

The p-value (or observed significance level) is characteristic of observed data indicative of the degree of refutation of the null hypothesis. If p is considered sufficiently low, we reject the null hypothesis, considering that we were able to show its fallacy and the result is declared significant (Poitevineau, 1998; Nickerson, 2000). The p-value itself is not without its critics, however. In particular, it depends on sample size: the larger the sample, the smaller the p-values, all things being equal. It becomes difficult to distinguish what is due to the magnitude of the effect tested (effect size) and what is due to the effect size of the sample (sample size) in any given p value. In all cases, whether the level of significance is α or p-value, the position of the cursor that determines the boundary between what is statistically significant and what is not will still be an arbitrary position.

Effect size

Effect size measures the amplitude or strength of the relationship between two or more variables in the population. For example, a 20% increase in premiums paid to employees may result in a 30% increase in their productivity. Tests have often neglected the effect size (Cohen, 1988, 1994; Schwab & Starbuck, 2009). However, as Poitevineau (1998) pointed out, a statistically significant result is an indication of the existence of the supposed effect, while a non statistically significant result is a statement of ignorance. Going ahead on the sole basis of the test would wrongly equate statistical significance and magnitude of the effect. In fact, the researcher is primarily interested in the magnitude and precision of effects. Let's consider, for example, the research hypothesis that training increases employee productivity. Here, the real issue is not so much whether training has an effect on productivity but rather whether this effect is significant enough to justify investing in a training plan. The use of NHST might lead to formulating the null hypothesis that training has no effect and, if the null hypothesis were to be rejected, to interpret the result as proof of the usefulness of training plans. Naturally, such a conclusion is not strictly justified by the outcome of NHST. In sum, the relevant questions remain the following: is the difference means trivial, low, medium or high? Is this difference meaningful? Is it significant enough to be included in a larger model? These important research questions are not all addressed by the NHST because they go beyond their scope.

Social scientists often have great difficulty in defining what is substantively significant. In other words, they have difficulty determining at what point a difference, proportion, correlation coefficient or regression, etc. becomes important, interesting or relevant. Although there are no absolute rules for interpreting the magnitude of effects, guidelines have been developed in different research fields. In this regard, the work by Cohen (1988) remains an essential benchmark. This author defined conventions that, over the years, have become virtually universal standards (small effect = 0.20, average effect = 0.50, large effect = 0.80).

Yet a careful reading of the book leads us to discover that the author warns against this situation and requires researchers to define the size of the effect for themselves. “The values chosen had no more reliable a basis than my own intuition. They were offered as conventions because they were needed in a research climate characterized by a neglect of attention to issues of magnitude” (Cohen, 1988: 532)

Technical problems: frequent errors in use

Beyond their intrinsic statistical defects, NHST are criticized for being the source of frequent errors committed by researchers, although it could be argued that the researchers themselves are most to blame for that. Below, we set out the main errors encountered (Sawyer & Peter, 1983; Poitevineau, 2004):

Reversal of conditions

The first and most typical error is to consider the p-value (observed significance level) or α (the significance level) as a probability of the null hypothesis (Pr (H0lData)) and not conditional on it (Pr (Datal H0). Many authors (Cohen, 1994; Poitevineau, 1998; Gill, 1999) warn against the following false claims:

- “the probability that the null hypothesis is true is p (or α)”;
- “the probability that the results are due to chance alone is p (or α)”;
- “the probability that the alternative hypothesis is true is 1-p (or 1-α)”.

The p-value (observed significance level) is calculated by assuming that the null hypothesis is true, which means that any observed difference is solely due to chance. The p-value is then used to determine the rejection or non-rejection of the postulated null hypothesis. For example, a test of difference in scores between two samples shows a difference of 10 points, which is associated with a p-value of 3%. The p-value of 3% means that if the null hypothesis is true (that is, if people actually have the same mean score) then there are 3 in 100 chances of observing a score difference greater than or equal to 10 points, and 97 in 100 chances of obtaining a score difference less than or equal to 10 points between two arbitrary samples from each of the two populations. The p value is therefore clearly dependent on the data probabilities (the difference in observed scores) and not on the probability of the null hypothesis.

Errors of interpretation may be explained by the gap between what researchers expect and what the tests provide (Cohen, 1994). Researchers use NHST to decide whether their findings support or refute their hypothesis. But tests indicate the probability of obtaining the observed results (given that the null hypothesis is true) and not the probability of the null hypothesis (given the data). We have already noted that the values of these two probabilities could differ significantly, even if it is true that the smaller that p is, the greater the evidence against the null hypothesis. In very general terms, a statistically significant finding is a result that occurs very rarely when the null hypothesis is true (Sawyer & Peter, 1983).

There are, however, similar errors on confidence intervals. Often, researchers report a confidence interval of around 95% [X, Y]: the parameter p has a 0.95 probability of being in the range [X, Y]. This natural

interpretation is nonetheless erroneous. The parameters are generally unknown but they have a fixed rather than random value. The event «X <p <Y» is true or false (because p is fixed), and we cannot assign a probability to it (other than 1 or 0). The correct interpretation of the 95% confidence interval is as follows: «95% of the intervals calculated over all possible samples (those that can be drawn in the population) contain the true value p. Each interval has a particular probability 0 or 1 to contain the true value. Here, this is not the random parameter but the limits of the confidence interval which vary from one sample to another.

Probability of replicating the results

A second mistake is to consider (1-p) or (1-α) as the probability of replicating the observed outcome (Falk & Greenbaum, 1995; Poitevineau, 1998; Gill, 1999). Traditional theories of statistical tests provide no indication of the likelihood of replicating the observed result. This replicability can be seen in two ways:

- it may only concern the significance of the result, which is the most common case and gives statements such as: "the probability that a replication is significant is 1-p (or 1-α)";
- it may also concern the very value of the effect: "there is a 95% chance of observing a similar result in future work."

These statements are of course wrong, even if it is true that the higher the observed significance level p, the more replicability is ensured. In fact, the replicability of results crucially depends on the controllability of the relevant variables, as is the case of the experiments. For example, a study may have demonstrated a statistically significant positive relationship between the adoption of a specific training technique and staff productivity. This result may not be repeated in other studies simply because a variable as important as the quality of the trainer was not taken into consideration, when it was in fact the real cause of improved productivity, more than the training technique itself. Clearly, nothing in the logic of NHST allows for the interpretation of a statistically significant result as a direct indication of the likelihood of replicating the findings (Carver, 1978). Another argument put forward by Armstrong (2007b) inspires the following illustration: suppose there is a correlation of 0.3 between the adoption of a training plan and an increase in staff productivity. If fifty replication attempts were conducted using tests with a power of 50% (a reasonable value in social science), then about half the replication attempts would conclude that there is correlation and the other half a lack of correlation at a significance level α of 5%. In this case, an NHST conducted on a meta-analysis of 50 replications would lead us to wrongly conclude that there is no statistically significant correlation between the adoption of a specific training technique and staff productivity.

Statistical significance and substantive significance

The third mistake is to confuse statistical significance with substantial significance (Sawyer & Peter, 1983; Gliner, Morgan, Leech, & Harmon, 2001). This leads to the belief that the more a result is statistically significant, the more scientifically interesting it is, and/or the larger the

effect with respect to the population. The arguments previously devoted to the magnitude of the effect (effect size) and the effect of sample size showed the difference between the two significances. We should remember, following Carver (1978), that a statistically significant result is literally a result whose probability of occurrence is low if the null hypothesis is true. Now, as the statistical significance depends on the size of the sample, trivial differences are often interpreted as significant when the sample size is very large. For example, suppose that employee productivity is measured on a scale of 0-100 in two samples, one with employees who attended a specific training course and the other with employees who did not. A difference of one hundredth of a point is found between the productivity mean scores of the two samples. A hundredth of a point difference in a performance score between 0 and 100 is clearly trivial. However, this trivial difference will be statistically significant if each sample consists of thousands of employees. Conversely, a substantial difference (e.g. multiple points) can be statistically insignificant if the samples are small enough in size (just a few individuals). In the latter case, Carver (1978) suggests attempting to check if replication again gives an effect of comparable intensity. Ultimately, it is up to the researcher who has thought about his or her research hypothesis to identify what makes sense (i.e. whether, in terms of meaning, a nonzero constant is not a better choice for the null hypothesis than the unlikely strictly zero mean scores difference). In any event, and on a broader level, several authors (Carver, 1993; Thompson 1996) emphasize that the expression “statistically significant” should never be replaced simply by «significant.»

Accepting the null hypothesis

A fourth mistake is to conclude that the null hypothesis is true if the result is not significant. For example, a researcher examines the research hypothesis that «the establishment of a training plan increases staff performance. A mean difference test of performance scores leads her to observe a statistically insignificant result: there is no evidence of a statistically significant difference between the performance scores of groups with or without a training plan. The researcher must conclude that she cannot reject the null hypothesis and should certainly not say that she «accepts» the null hypothesis. For example, she should not say: “the establishment of a training plan does not increase staff performance.” This would be equivalent to accepting the null hypothesis and concluding there is a lack of effect by way of inference to the general population, instead of merely stating a descriptive comment. Similarly, the phrase “since the value 0 is included in the interval, we cannot reject the null hypothesis that the two sets of values have the same mean” is correct because it is a description of the samples. However, we should not say: “the distribution of stock options does not alter executives’ loyalty” because this is an inference. In fact, Cohen (1994) showed that proving the null hypothesis is a logical impossibility in the context of NHST. Several other authors have also warned against such an error (Gill, 1999; Krueger, 2001).

The previous errors frequently made by researchers may sound like ad-

ditional criticism of NHST: a method that produces so many errors in its application, even among advanced users, can hardly plead innocent. But NHST have yet a third type of problems, philosophical this time.

Philosophical problems: the limits of inference

The logic of NHST is based on inference (Fisher, 1942; Krueger, 2001; Morgan, 2003). The precursors of inference such as David Hume in the mid 18th century and Karl Pearson in the early 20th century argue that future events can be predicted from the sequences and frequencies of past events (Alexander, 1972; MacNabb, 1972; Morgan, 2003). According to Hume and Pearson, causes and effects can only be justified as an extended series of coincidences that we begin to associate with anticipation (Black, 1972). One of the philosophical implications of this conceptualization is that, although it is possible to prove that something is wrong (because of lack of coincidence), it becomes impossible to prove that something is true (Howell, 1997; Krueger, 2001; Morgan, 2003).

Fisher, like Hume and Pearson, believed that inference is basically the only process enabling the discovery of new knowledge (Fisher, 1942). Therefore, the objective of Fisher's inference system is to test and, more specifically, to refute a hypothesis that a particular treatment has led to differences between samples (Mulaik, et al., 1997; Morgan, 2003). Central to inference, the assumption that the future resembles the past is not based on any argument but is simply derived from the habit whereby we are determined to expect the same set of things in the future to which we have been accustomed (Hume, 1978). And as Morgan (2003) points out, analysis of experimental data leads to inferences about the likelihood of future events: when the differences between conditions are unlikely under the null hypothesis, researchers attribute this difference to the stability of underlying causes and therefore expect to observe the same differences in similar circumstances at another time.

The general approach is to make an assumption, observe a real phenomenon and then assess the compatibility of the base case with the real phenomenon. More specifically, the reasoning corresponds to the following syllogism (Gill, 1999):

1. If A then B;
2. Not B is observed;
3. Therefore not A

For NHST, this reasoning becomes:

1. If H0 is true then the data will have an expected particular pattern;
2. The data did not have the expected particular pattern;
3. Thus H0 is false.

The main problem is that in the practical application of this formal logic to NHST, we attach certain assertions to statements of probability. Indeed, the argument becomes:

1. If A then B is highly likely
2. Not B is observed;
3. Thus A is highly unlikely

For NHST, the reasoning becomes:

1. If H0 is true then the data will most likely be of a particular pattern;
2. The data did not have the expected particular pattern;
3. Thus H0 is very probably false.

At first glance, this logic seems plausible. Yet it is wrong to assert that the presence of data that are atypical or improbable under a given assumption implies that the assumption is false. It may simply be that a rare or unlikely phenomenon occurred (Cohen, 1994; Gill, 1999). Consider the following example:

1. If a person speaks French, she is probably not a member of the French government;
2. The person is a member of the French government;
3. So she very probably does not speak French.

This example, with its absurd conclusion, clearly shows the limits of statistical inference based on NHST. Rejecting the null hypothesis in the context of an NHST merely suggests that the results should not be attributed to chance. In other words, it suggests that "there is not nothing" in the words of Dawes (1991: 252). This probabilistic inference amounts to proof by contradiction (modus tollens). If the null hypothesis is true, the existence of ordered data is unlikely. If the data seems unlikely, then the null hypothesis is probably false. If the null hypothesis is false, then something fundamental, and other than chance, is probably taking place (Chow, 1998; Morgan, 2003). The main problem with this chain of inference is that syllogisms are not valid when applied to inference. Three criticisms are primarily advanced (Morgan, 2003). First, any ad hoc hypothesis is false, and at the limit, no data is needed to reject it (Tukey, 1991; Morgan, 2003). Therefore, the purpose of NHST should be anything other than mere rejection of the null hypotheses. Second, even when assuming that a hypothesis is true, the probability of data confirming a hypothesis does not necessarily mean that the reverse is true. For example, the probability that a firm detaining exactly ten key success factors will be successful is not the same as the probability that a successful firm detains exactly ten key success factors. The first probability is in any case very strong and the second in any case very low. No contradiction, as unlikely as it is, can refute anything if the premises are uncertain. Thirdly, NHST are of no special help when it comes to assessing the possibility of replicating the results in the future (Carver, 1978; Morgan, 2003).

Hume (1978) noted that inference cannot be validated other than by inference itself. Inference from a sample of any size cannot provide certain knowledge about the population's characteristics. However, because our inferences may have worked in the past, we hope they will continue in the future. This in itself is an inference that can be justified only by other inferences, and so on (Krueger, 2001; Morgan, 2003). Empirical research must either accept this act of faith or be broken. Because knowledge "must include reliable predictions" (Reichenbach, 1951: 89), we "act as if we have solved the problem of induction" (Dawes, 1997: 387).

Interestingly, it is a reflection on NHST that may have led to the (re)discovery that the researcher's work, in which the process of inference is central, fundamentally requires an act of faith (Hume, 1978; Reichenbach, 1951 Dawes, 1997; Krueger, 2001). Indeed, the use of NHST and, more generally, concern for statistical generalization, are essentially relevant for positivist researchers. Non positivist researchers might be interested in making sense of a certain class of phenomena without attempting to identify general laws that would govern them. Therefore, the principal focus of the debate on NHST remains to show that they are an ordinary tool for the development of inference. All of this refers to the ideal qualities of (management) researchers and their ability to strike a balance (always fragile since dynamic) between on the one hand, daring and determination to be creative and inventive and, on the other hand, caution, hesitation, doubt, humility and respect for the data and the outside world that inspires and entertains the theories and hypotheses of the (management) researcher.

HOW TO BEST USE STATISTICAL SIGNIFICANCE TESTS

IDENTIFYING THE CAUSES OF PERSISTENT PROBLEMS

In spite of the three types of problem posed by NHST (statistical, technical, philosophical), their popularity remains high among researchers (Armstrong, 2007b; Levine, Weber, Hullett, Hee Sun Park, & Lindsey, 2008). Despite the many persistent criticisms their use is subject to, NHST are conventionally accepted as proof of the validity of findings and are a gold standard for the publication of research findings. It is as if we were in the presence of a practice that is theoretically and methodologically questionable but sociologically adapted, a tool used for evil because its use is particularly misleading but that nevertheless enjoys an aura hitherto untouched. Poitevineau (1998, 2004) summarizes the main reasons for this apparent paradox:

- The ambiguity of terminology: NHST are "significance tests", which refers to "significant", something that gives meaning, which is of importance, etc. Consequently, the confusion between statistical and substantial significance is induced.
- Objectivity: researchers look for formalized and objective methods that will enable them to know if a data set contains random or systematic variations. And they consider it important not to have to rely on their own intuition and subjectivity in determining the proportion of random and systematic effects in the data. Therefore, NHST provide the researchers' conclusions with that sense of objectivity that is in their vital interest.
- Scientific nature: in disciplines such as management that suffer more or less from a complex of being of a non-scientific nature, at least in relation to 'harder' sciences, the mathematical apparatus and formalism of NHST provide a cheap scientific ve-

neer. In addition, the rigor of mathematics and its supposed aura spread across research, ensuring its validity de facto.

• Reinforcement by Karl Popper: NHST offer a great resemblance to Popper's idea that the demarcation between scientific and unscientific statements is made on the basis of their falsifiability or refutability. A scientific hypothesis is a hypothesis that can be empirically 'tested'. NHST theory has benefited from the success of Popper's ideas.

• Guaranteed comfort and economy: NHST provide some comfort to their users. With their power to declare an effect as "significant", NHST are seen as a solution, relieving the researcher from the task of interpretation, as if statistical significance was sufficient in itself.

Everything leads us to believe that the continued success of NHST is due to a tremendous misunderstanding: a semblance of objectivity and scientific nature and an illusion of adequacy to researchers' needs permitted by the ignorance that most researchers have about the nature and conditions of use of such NHST. However, criticisms (which are not new) are slowly beginning to have their effect. The trigger was not a sudden awakening or awareness by researchers using NHST but rather institutions like the American Psychological Association or the editorial boards of scientific journals prescribing new standards for publication. Essentially, the results of traditional statistical analysis should be completed, beyond the observed significance levels alone or p-values for the systematic inclusion of indicators of the magnitude of effects and their interval estimates. In this vein, what, more generally, are the main avenues for improvement?

SOME AREAS FOR IMPROVEMENT

Several areas for improvement are possible. A practical approach is to start with the recommendations by the Task Force as set out by the Office of Scientific Affairs of the American Psychological Association (APA) to study the role of NHST in psychological research (APA, 1996). We will then explore complementary areas for improvement.

Recommendations of the American Psychological Association

• Hypothesis testing: it is difficult to imagine a single situation where a binary decision of acceptance/rejection would be preferable to reporting the p-values or, better still, a confidence interval. Moreover, one should never use the unfortunate expression "accept the null hypothesis."

• Intervals: intervals should be provided for any effect size concerning the main results. Such intervals should be provided for correlations and association or variation indices whenever possible.

• Effect sizes: always present effect sizes for the raw results. If the units of measurement are meaningful practice (e.g. number of cigarettes smoked per day), prefer a non-standardized measure (regression coefficient or mean difference) to a standardized measure.

• Power and sample size: always provide information on the size of the sample and the process that led to the choice of such a size as well as explicit assumptions about the magnitude of effects, sampling and measurement of variables and the analytical procedures used for the calculation of power. Insofar as the power calculation is more meaningful when done before the collection and review of data, it is important to show how estimates of the magnitude of effects were derived from previous research and theory to remove the suspicion that the data were derived uniquely from the current study or, still worse, were constructed to justify a given sample.

Complementary statistical methods

Many statisticians have been pleading for some time for statistical methods other than NHST (Gill, 1999; Nickerson, 2000). Among these alternatives to NHST statistical methods are the likelihood methods and Bayesian methods (Poitevineau, 1998).

• The likelihood methods: in the simple case of two ad hoc hypotheses H0 and H1, the likelihood ratio method involves calculating the probability density of the observed statistics (x) under H0 and under H1, e.g., f(xlH0)/f(xlH1). This ratio represents the likelihood of one hypothesis over another on the basis of observed results. One may possibly retain H0 or H1 depending on whether this ratio is greater or less than an arbitrarily chosen constant (one, for example, if no hypothesis is favored). The likelihood ratio method has the advantage of using neither priors nor non observed elements. While the likelihood ratio allows the strength of empirical evidence between two ad hoc point hypotheses to be evaluated, it is unfortunately very rare in practice that a researcher is confronted with such a simple case.

• Bayesian Methods: used as a method of statistical inference, the Bayesian approach involves using Bayes' theorem to calculate the posterior distribution of the parameter we are interested in, based on:

- o observed data;
- o a sampling model;
- o a priori probabilities of the parameter (priors).

Several authors have advocated the replacement of conventional NHST by a Bayesian approach (Edwards, Lindman, & Savage, 1963; Rouanet, 1996). Unlike conventional NHST, the Bayesian approach directly affects the probability of the truth of the research hypothesis (Bakan, 1966; Carver, 1978). The Bayesian approach has been, and still is, widely criticized as too subjective an approach because it requires a priori probabilities to be specified. However, the weight of prior distribution in the posterior distribution diminishes as the mass of data increases. Thus, two researchers using different priors will arrive at similar conclusions if the data are sufficient. It is also recommended to vary the prior distributions (optimistic, neutral, and pessimistic positions) and to analyze the sensitivity of the results. In fact, Bayesian methods seem to present enough advantages to emerge as genuine

challengers for NHST. There are many examples of the use of Bayesian methods in management science, particularly in finance (Corless, 1972; Holt & Morrow, 1992; Sarkar & Sriram, 2001) and marketing (Roberts, 1963, Levitt 1972). The research by Albert, Grenier, Denis, & Rousseau (2008) devoted to the study of food risk is also highly relevant to management researchers. Furthermore, an increasing amount of statistical software now incorporates Bayesian analysis modules (for example, SPSS with Amos or, more recently, MPLUS).

Beyond understanding the recommendations of institutions such as the American Psychological Association or considering new methods of statistical inference (e.g. Bayesian methods), a third means of improvement which seems by far the most important concerns the researcher's attitude.

Back to a researcher position

The main challenge is to stick to some of the basic qualities of researchers such as critical thinking, alertness, doubt, boldness, creativity, strength of will, etc.

Various reasons, including sociological, historical, cognitive, emotional reasons, etc. can sometimes lead researchers to lack distance and a critical perspective with regard to their work environment, in particular vis-à-vis available research tools. The type of research training received (school of thought, profile of teachers and peers), the dominant paradigms in the structural environment (research centers, academic associations, etc.), as well as preferences or personal skills, will structure and shape researchers beliefs and attitudes to a large extent. These aspects can naturally encourage imitation and inhibit critical thinking in research methodology. However, the best research that can produce the most interesting results undoubtedly requires going beyond basic mimicry and the routine use of commonly used research methods and tools at some point in time. Turning specifically to NHST, we realize that they are certainly an important factor in the selection of articles submitted for publication nowadays, in the sense that an insignificant result is generally still very unlikely to be published. The low number of insignificant results published may well be the result of a deliberate editorial policy, or selection or censorship made by the researchers themselves. In any case, there is a very low rate of publication of such insignificant results. This can lead to catastrophic consequences. Let's consider for a moment the following scenario: several researchers are testing, independently of one another, the same null hypothesis H0 which is true. About 5% of them find a significant result (rejecting H0 at 5%) and are virtually the only ones able to publish, thus suggesting the reality of the phenomenon under study (reject H0). We would therefore be confronted with only spurious findings in the literature. And attempts at replication by audacious researchers would only worsen the situation: only statistically significant results would be selected and published in the future. What assurance do we have of not finding ourselves in such a scenario when we produce a literature review? Virtually none. This again illustrates the need for due diligence, boldness and a critical perspective on the part of the researcher. And these qualities become even more vital

when the tools are more sophisticated, numerous and readily available. Clearly, this individual requirement to promote the fundamental qualities of researchers should be accompanied by collective action to promote the publication of non statistically significant results, thereby reducing a serious threat to the researchers' professional environment. Within this perspective, it is interesting to note that, in scientific disciplines other than management, some journals already encourage the publication of non scientific insignificant results such as, for example, variants of the "Journal of Negative Results." Other scientific journals, including several journals from the International Committee of Medical Journal Editors, require an experiment to be registered before being undertaken to avoid self-censorship.

CONCLUSION

This article has sought to raise management researchers' attention to the dangers of the indiscriminate use of NHST. It builds on a series of publications that have nurtured and continue to fuel criticism of NHST. These publications cover virtually all fields: statistics (Berkson, 1942), psychology (Hunter, 1997), sociology (Selvin, 1957), marketing (Sawyer & Peter, 1983), accounting (Lindsay, 1995), political science (Gill, 1999), science education (Morgan (2003), psychiatry (Gliner, et al., 2001), forecasting (Armstrong, 2007a), ecology (Anderson, Burnham & Thompson, 2000; Gibbons, Crout, & Healey, 2007), meteorology (Nicholls, 2001), communication (Levine, et al., 2008), etc.

Criticism of NHST gained fresh momentum from the mid-1990s, and it gradually spread from statistics to psychology before affecting virtually all disciplines, with the notable exception of management which is just beginning to address the issue (Mbengue, 2007; Schwab & Starbuck, 2009). This was the main purpose of the present article: to inform the community of researchers in management of the existence of this criticism of NHST, present the content and issues in detail (the dangers of indiscriminate use of NHST), and offer recommendations for improving the process of testing hypotheses and, more generally, the process of (statistical) inference in (management) research. Compared to the paper by Schwab & Starbuck (2009), our text engages a wider literature, offers an in-depth discussion of inference, organizes the issues posed by NHST into three types (statistical, technical and philosophical), and provides several management-related examples and many concrete recommendations organized into three categories (respect of APA recommendations, the use of complementary or alternative statistical methods to NHST, and a return to fundamental researcher qualities).

There is general agreement on the dangers of NHST use. The first danger for researchers using NHST is to ignore their instructions, that is, their conditions of use. This danger is particularly threatening given the increasing availability of statistical software. Another danger for the researcher is to hide behind the scientific image of statistical tests, to yield to them and to the apparent comfort related to their use, hence abdicating responsibility. It is the researcher who must choose whether or not to test, what they test and by what means. But more importantly, researchers must bear in mind that NHST are just a tool within a mechanism and a research process: this research process begins before an eventual test, continues while testing and continues after the test. As for the test itself, it is only a tool and, as such, it is only useful if we use it wisely. From this point of view, the recurring questions about the usefulness of NHST provide a good incentive and a valuable safeguard for the exercise of sound, well-grounded research.

We began this article noting that NHST were at the heart of inferential statistics and, consequently, the process of inference (Krueger, 2001; Morgan, 2003). We also showed that no statistical method has been criticized so much while remaining eminently popular and widely misused (Krueger, 2001; Armstrong, 2007a, 2007b, Levine, et al., 2008).

The question posed in this article was whether, ultimately, it was necessary or not to burn NHST.

The analysis of criticism of NHST and awareness of alternative possibilities could result in an affirmative answer: after all, it does not appear that the abolition of NHST would jeopardize the process of inference, much less scientific research activity. However, many NHST defects appear to be related to their inappropriate use, calling into question NHST users rather then the tool itself. In this sense, the killing of NHST would be akin to a sentence for the least excessive. Of course, we can obviously blame a tool for not being sufficiently easy to use, which could ultimately lead to an intermediary verdict between capital punishment and acquittal.

In fact, the question of whether to burn NHST or not is inherently more interesting than any answer (positive, negative, intermediate) that might be given. Indeed, such a question basically refers to the place of inference in scientific research. But NHST do not have the monopoly of statistical inference, much less of the process of inference in general. Therefore, their criticism can hardly be discussed apart from a more general reflection on the nature of inference and its status in the work of researchers, as outlined in this article.

Finally, our paper is less iconoclastic than its title might suggest. It can be read as an answer to the question of how to better use NHST. Upon analysis, it is not really necessary to ban the use of NHST. Our article simply warns against indiscriminate and routine use of NHST. While the use of this tool has been abundantly criticized in most scientific disciplines, this unfortunately has not been the case in the field of management. Most of our text aimed to alert management researchers about the three types of problems (statistical, technical, philosophical) posed by NHST. The article therefore calls for more rigorous, more conscious, more reflective and more critical use of these tests, while suggesting the possible use of other statistical methods, namely the Bayesian methods and point or interval methods of estimation.

Of course, our study has not exhausted all the issues raised by the use of NHST. Several avenues for future research are open. A first avenue would be to conduct a large quantitative survey on management researchers' practices. Certainly, there is little reason to doubt that the field of management would not differ from any other discipline in which investigations to date have produced consistent results regarding the widespread prevalence of errors in NHST use. Certainly, too, many qualitative elements suggest to us that few management researchers know about the existence, let alone the content, of the criticism regarding NHST. However, only a quantitative survey would reveal the exact extent and nature of the evil or risk. For example, what are the most common mistakes in the management research community and in what circumstances is the researcher most exposed to them? Such an investigation could include published articles, which has been the main approach adopted so far in the work conducted in other disciplines, but also the practices and knowledge of researchers measured through interviews or questionnaires. Regarding the second method,

the only exception, to our knowledge, is the study by Mittag & Thompson (2000). The latter method of investigation seems very important in view of the increasing role of the gray literature, with the development of Internet and academic conferences (with or without proceedings) in the dissemination of good (or bad) research practices. A second line of research could be to conduct a meta-analysis to compare the diachronic and/or cross-sectional, that is between disciplines in a mode similar to previous work (which dates back several decades already) by Morrison & Henkel (1970). All these large quantitative surveys could be usefully combined with fine quality studies in order to achieve accurate diagnoses that might lead to avenues for effective therapies. We hope that many researchers will explore these important issues. More importantly, we hope they will challenge their fundamental qualities as researchers, in other words critical thinking, the refusal of mimesis, the cult of doubt, creativity and perseverance...!

Ababacat MBENGUE is professor at the University of Reims and at Reims Management School. His research interests include strategic management and the reinforcement of organizational capacities, knowledge management and methodology. He has been a visiting professor at Wharton (Snider Enterpreneurial Center) and at the University of Orel (Russia).

REFERENCES

■ Albert, I., Grenier, E., Denis, J. B., & Rousseau, J. (2008). Quantitative Risk Assessment from Farm to Fork and Beyond: A Global Bayesian Approach Concerning Food-Borne Diseases. *Risk Analysis, 28*(2), 557-571.

■ Alexander, P. (1972). Karl Pearson. The encyclopedia of philosophy. *New York: Macmillian, 6,* 68-69.

■ Anderson, D. R, Burnham, K. P., & Thompson, W. L. (2000). Null Hypothesis Testing: Problems, Prevalence, and an Alternative. *Journal of Wildlife Management, 64*(4), 912-923.

■ APA (1996). Task Force on Statistical Inference Report. Washington, DC: American Psychological Association.

■ Armstrong, J. S. (2007a). Significance tests harm progress in forecasting. *International Journal of Forecasting, 23*(2), 321-327.

■ Armstrong, J. S. (2007b). Statistical significance tests are unnecessary even when properly done and properly interpreted: Reply to commentaries. *International Journal of Forecasting, 23*(2), 335-336.

■ Baillargeon, G., & Rainville, J. (1978). *Statistique appliquée* (Tome 2, 6e édition). Trois-Rivières: Les Éditions SMG.

■ Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423-437.

■ Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37*, 325-335.

■ Black, M. (1972). Induction. The encyclopedia of philosophy. *New York: MacMillan, 4*, 169-181.

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

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

■ Chow, S. L. (1998). Statistical significance: Rationale, validity and utility. *Behavioral and Brain Sciences, 21*, 169-240.

■ Cohen, J. (1988). *Statistical power analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

■ Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 55-159.

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

■ Corless, J. C. (1972). Assessing prior distributions for applying Bayesian statistics in auditing. *Accounting Review, 47*, 556-566.

■ Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics, 29*, 357-372.

■ Dawes, R. M. (1991). Probabilistic versus causal thinking. In D. Cicchetti & W. M. Grove (Eds.), Thinking clearly about psychology: Vol. 1. Matters of public interest: Essays in honor of Paul Everett Meehl (pp. 235-264). Minneapolis: University of Minnesota Press.

■ Dawes, R. M. (1997). *Qualitative consistency masquerading as quantitative fit.* In M. L. Dalla Chiara, D. Kees, D. Mundici & J. van Bentheim (Eds.), Structures and norms in science (pp. 387-394). Dordrecht, the Netherlands: Kluwer Academic.

■ Dodge, Y. (1993). *Statistique : Dictionnaire encyclopédique.* Paris : Dunod.

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

■ Falk, R., & Greenbaum, C. W. (1995). Significance Tests Die Hard. *Theory and Psychology, 5,* 396-400.

■ Firestone, W. A. (1993). Alternative arguments for generalizing from data as applied to qualitative research. *Educational Researcher, 22*(4), 16-23.

■ Finch, S., Cumming, G., & Thomason, N. (2001).
Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement, 61,* 181-210.

■ Fisher, R. A. (1942).
The design of experiments (3rd ed.). London: Oliver & Boyd.

■ Gibbons, J. M., Crout, N. M., & Healey, J. R. (2007).
What role should null-hypothesis significance tests have in statistical education and hypothesis falsification? *Trends in Ecology & Evolution, 22*(9), 445-446.

■ Gill, J. (1999).
The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly, 52(*3), 647-674.

■ Gliner, J. A., Morgan, G. A., Leech, N. L., & Harmon, R. J. (2001).
Problems with null hypothesis significance tests. Journal of the American *Academy of Child and Adolescent Psychiatry, 40,* 250-252.

■ Hagen, R. L. (1997).
In praise of the null hypothesis statistical test. *American Psychologist, 52*, 15-24.

■ Harlow, L. L. (1997).
Significance testing introduction and overview. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 1-21). Mahwah, NJ: Lawrence Erlbaum.

■ Holt, D. L., & Morrow, P.C. (1992).
Risk assessment judgments of auditors and bank lenders: A comparative analysis of conformance to Bayes' theorem. Accounting, *Organizations and Society, 17*(6), 549-559.

■ Howell, D. C. (1997).
Statistical methods for psychology (4th ed.). Belmont, CA: Duxbury.

■ Hubbard, R., & Bayarri, M. J. (2003).
Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician, 57*, 171-178.

■ Hume, D. (1978).
A treatise of human nature. Glasgow, Scotland: William Collins (original datant de 1739).

■ Hunter, J. (1997).
Needed: A ban on the significance test. *Psychological Science, 8,* 3-7.

■ Kanji, G. K. (1993).
100 Structural tests. Thousand Oaks: Sage.

■ Kennedy, M. M. (1979).
Generalizing from single case studies. *Evaluation Quarterly, 3*(4), 661-678.

■ Krueger, J. (2001).
Null Hypothesis Significance Testing. *American Psychologist, 56*(1), 16-26.

■ Levine, T., Weber, R., Hullett, C., Park, H. S., & Lindsey, L. (2008).
A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research. *Human Communication Research, 34*(2), 171-187.

■ Levitt, T. (1972).
Industrial Purchasing Behavior: A Bayesian Reanalysis. *Journal of Business Administration, 4,* 79-81.

■ Lindsay, R. M. (1995).
Reconsidering the Status of Tests of Significance: An Alternative Criterion of Adequacy. *Accounting, Organizations and Society, 20*, 35-53.

■ MacNabb, D. (1972).
David Hume. *The encyclopedia of philosophy.* New York: MacMillan, 4, 74-90.

■ Martinez-Pons, M. (1999).
Statistics in modern research: Applications in the social sciences and education. New York: New York University.

■ Mbengue, A. 2007.
Tests statistiques de signification. In R.-A. Thietart & coll. (Eds.), Méthodes de recherche en management (pp. 297-349). Paris: Dunod.

■ Meehl, P. E. (1990).
Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 56* (Suppl. 1), 195-244.

■ Mittag, K. C., & Thompson, B. (2000).
A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*(4), 14-20.

■ Morgan, P. L. (2003).
Null Hypothesis Significance Testing: Philosophical and Practical Considerations of a Statistical Controversy. *Exceptionality, 11*(4), 209-221.

■ Morrison, D. E, & Henkel, R. E. (Eds.) (1970).
The Significance Test Controversy. Chicago: Aldine.

■ Mulaik, S.A., Raju, N.S., & Harshman, R.A. (1997).
There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 65-116). Mahwah, NJ: Lawrence Erlbaum.

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

■ Nicholls, N. (2001).
The insignificance of significance testing. *Bulletin of the American Meteorological Society, 82,* 981-986.

■ Nickerson, R. S. (2000).
Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy. *Psychological Methods, 5*(2), 241-301.

■ Nunnally, J. (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 behavioral sciences. New York: Wiley.

■ Poitevineau, J. (1998).
Méthodologie de l'analyse des données expérimentales : étude de la pratique des tests statistiques chez les chercheurs en psychologie, approches normative, prescriptive et descriptive. Thèse de Doctorat, Université de Rouen.

■ Poitevineau, J. (2004).
L'usage des tests statistiques par les chercheurs en psychologie : aspects normatif, descriptif et prescriptif. *Mathématiques et Sciences Humaines, 167*, 5-25.

■ Reichenbach, H. (1951).
The rise of scientific philosophy. Berkeley: University of California Press.

■ Roberts, H. V. (1963).
Bayesian Statistics in Marketing. *Journal of Marketing, 27,* 1-4.

■ Rozeboom, W. W. (1960).
The Fallacy of The Null-Hypothesis Significance Test. *Psychological Bulletin, 57*(5), 416-428.

■ Rouanet, H. (1996).
Bayesian methods for assessing importance of effects. *Psychological Bulletin, 119*, 149-158.

■ Sarkar, S., & Sriram, R. S. (2001).
Bayesian Models for Early Warning of Bank Failures. *Management Science, 47*(11), 1457-1475.

■ Sawyer, A. G., & Peter, J. P. (1983).
The Significance of Statistical Significance Tests in Marketing Research. *Journal of Marketing Research, 20*(2), 122-133.

■ Schmidt, F. L., & Hunter, J. E. (2002).
Are there benefits from NHST? *American Psychologist, 57*, 65-66.

■ Schwab, A., & Starbuck, W. H. (2009).
Null-hypothesis significance tests in behavioral and management research: we can do better. *Research Methodology in Strategy and Management, 5*, 29-54.

■ Selvin, H. C. (1957).
A Critique of Tests of Significance in Survey Research. *American Sociological Review, 22*, 519-527.

■ Sincich, T. (1996).
Business statistics by example. Upper Saddle River, NJ: Prentice-Hall.

■ Smaling, A. (2003).
Inductive, analogical, and communicative generalization. *International Journal of Qualitative Methods, 2*(1), 1-31.

■ Thompson, B. (1996).
AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

■ Tukey, J. W. (1991).
The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

■ Wainer, H. (1999).
One cheer for null hypothesis significance testing. *Psychological Methods, 6*, 212-213.

■ Yin, R. (1984).
Case study research: Design and methods. Beverly Hills, CA: Sage.

■ Zikmund, W. G. (1994).
Business research methods. Orlando, Florida: The Dryden Press.

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