The Controversy over Null Hypothesis Significance Testing Revisited

Nekane Balluerka[1], Juana Gómez[2], and Dolores Hidalgo[3]

[1]University of the Basque Country, San Sebastian, [2]University of Barcelona, [3]University of Murcia, all Spain

Abstract. Null hypothesis significance testing (NHST) is one of the most widely used methods for testing hypotheses in psychological research. However, it has remained shrouded in controversy throughout the almost seventy years of its existence. The present article reviews both the main criticisms of the method as well as the alternatives which have been put forward to complement or replace it. It focuses basically on those alternatives whose use is recommended by the Task Force on Statistical Inference (TFSI) of the APA (Wilkinson and TFSI, 1999) in the interests of improving the working methods of researchers with respect to statistical analysis and data interpretation. In addition, the arguments used to reject each of the criticisms levelled against NHST are reviewed and the main problems with each of the alternatives are pointed out. It is concluded that rigorous research activity requires use of NHST in the appropriate context, the complementary use of other methods which provide information about aspects not addressed by NHST, and adherence to a series of recommendations which promote its rational use in psychological research.

Keywords: statistical significance, null hypothesis testing, psychological research, methodology

Null hypothesis significance testing (NHST) is a common practice in psychological research. According to Gigerenzer and Murray (1987) it became established as the main method of inductive inference between 1940 and 1955, a period which they refer to as marking the "inference revolution" in psychology. Both the study of Hubbard, Parsa and Luthy (1997), based on the *Journal of Applied Psychology*, and that of Hubbard and Ryan (2000), which considered a random sample of articles published in twelve journals of the American Psychological Association, illustrate that the use of *p* values in empirical research has grown rapidly since the 1950s; at that time 70% of studies based their analysis on this index whereas the figure since 1990 has risen to over 90%.

However, throughout its almost seventy years of existence NHST has been shrouded in controversy. Among the earliest criticisms of the logic and usefulness of NHST one of the most serious was that of Joseph Berkson in 1938. Since then, numerous authors have continued to highlight the problems associated with the method, critical papers being published in the 1960s (for example, Bakan, 1966; Cohen, 1962; Grant, 1962; Lykken, 1968; Meehl, 1967; Rozeboom, 1960), the 1970s (for example, Carver, 1978; Cronbach, 1975; Greenwald, 1975; Meehl, 1978; Morrison & Henkel, 1970; Tversky & Kahneman, 1971) and the 1980s (for example, Brewer, 1985; Cohen, 1988; Dar, 1987; Falk, 1986; Gigerenzer & Murray, 1987; Gigerenzer et al., 1989; Guttman, 1985; Huberty, 1987; Kupfersmid, 1988; Oakes, 1986; Rosnow & Rosenthal, 1989; Sedlmeier & Gigerenzer, 1989). Since 1990 the number of dissenting voices has increased still further (for example, Carver, 1993; Cohen, 1990, 1994; Dar, Serlin, & Omer, 1994; Falk & Greenbaum, 1995; Finch, Cumming, & Thomason, 2001; Gigerenzer, 1993; Harris, 1991; Hubbard, 1995; Hunter, 1997; Hunter & Schmidt, 1990; Kirk, 1996, 2001; Loftus, 1991, 1995, 1996; Meehl, 1990a, 1990b; Rossi, 1990, 1997; Shaver, 1993; Schmidt, 1992, 1996; Thompson, 1993, 1994, 1996, 1997; Tukey, 1991). However, a number of excellent publications defending the validity and usefulness of NHST have also appeared (for example, Abelson, 1995, 1997; Chow, 1987, 1988, 1989, 1991, 1996, 1998a,b; Cortina & Dunlap, 1997; Cox, 1977; Dixon, 1998; Frick, 1996; Hagen, 1997).

In light of the above the present article aims to examine the controversy surrounding use of NHST in psychological research. To this end, the paper begins by presenting the logic underlying the test and defines a series of important concepts associated with it. It then goes on to analyse both the main criticisms levelled at NHST as well as the alternatives which have been put forward to complement or replace it. The arguments in support of

DOI 10.1027/1614-1881.1.2.55

the validity and usefulness of NHST are then discussed. The article ends by proposing a way of overcoming the current controversy and of promoting the rational use of NHST in psychological research.

It is important to point out that unlike Nickerson's (2000) review of this issue, the present article addresses those alternatives to NHST which have been accepted by a large majority of authors who are critical of it, and whose use is recommended by the APA Task Force on Statistical Inference (TFSI) (Wilkinson & TFSI, 1999) in the interests of improving the working methods of researchers with respect to statistical analysis and data interpretation. Thus, for example, although the alternative based on Bayesian statistics is referred to at various points throughout the article this approach is not dealt with in a section of its own as, in our opinion, it generates as much controversy as NHST does. We therefore believe that it requires a detailed and separate review of its strengths and weaknesses, one which adopts a similar perspective to that used here with respect to NHST. In addition, the present article discusses a new alternative not referred to by Nickerson and which is becoming widely accepted by the scientific community, namely, the calculation of confidence intervals for effect sizes. Thus, although Nickerson's review is an excellent one the present study addresses various aspects not dealt with in his paper. Knowledge of these aspects may prove useful to all those researchers interested in improving their working methods with respect to statistical analysis and data interpretation.

The Concept of NHST

The null hypothesis significance test, proposed by Fisher (1925), may be used provided that the sample to which it is applied is random or representative of a population. The test involves establishing the acceptable probability (between 0 and 1) of committing an inferential error due to the sampling error inherent in the sample. The probability related to the decision to reject a null hypothesis H_0 (that is, a hypothesis which specifies that $\mu_1 = \mu_2 = \mu_3$, or $R^2 = 0$) when H_0 is true for the population is termed *alpha* (α) or *critical p*. This is the probability of making a wrong decision, known as Type I error. *Critical p* is established on the basis of a subjective judgment regarding the consequences which could result from committing a Type I error in a given study.

A second probability within the statistical significance test is *calculated p* or the *level of significance p of the empirical result*. This *p* expresses the probability (between 0 and 1) of obtaining the same or a higher sample statistic than that actually obtained, given a certain sample size and assuming that the sample was taken from a population in which H_0 is exactly true. The value of this probability depends, firstly, on the values of the true parameters of the population from which the sample is taken. Given that population parameters are unknown, it is assumed in the statistical significance test that they are correctly specified by H_0, that is, it is assumed that H_0 is exactly true in the population. A second factor influencing the computation of *calculated p* is sample size. Assuming that H_0 is exactly true, sample statistics will become less likely and, consequently, *calculated p* will be lower as sample sizes increase.

Bearing in mind these two probabilities, NHST tells us whether the calculated probability of our sample results is the same as or less than the acceptable limit *(critical p)* established for Type I error, that is, whether the results are due to sampling error, given our sample sizes and assuming that the sample has been taken from a population in which H_0 is exactly true. Therefore, when *calculated p* is the same as or less than *critical p* the decision is made to reject the H_0, a 'statistically significant' result thus being obtained; this implies that we consider our sample results to be unlikely under certain assumptions, including that regarding the veracity of H_0.

However, although it is easy to choose *critical p*, the computation of *calculated p* is tedious. Therefore the following steps are followed in testing H_0:

1. A H_0 is proposed regarding the value of a parameter in the population.
2. A sample distribution or a statistic is determined as an estimator of the population function.
3. A random sample of size *n* is obtained and the value of the estimator is calculated in this sample.
4. $p(D/H_0)$ is calculated, that is, the probability of obtaining results which are equally or more discrepant than those actually obtained (D), given a certain sample size and assuming that H_0 is exactly true in the population.
5. If $p(D/H_0)$, or the *calculated p*, is the same as or less than the *critical p* or α the H_0 is rejected. If, in contrast, the *calculated p* is higher than the *critical p*, the H_0 is provisionally accepted. The acceptance of the H_0 is tentative because Fisher's proposal is based on the Popperian approach to knowledge building, in which the falsity, but not the veracity, of theoretical formulations can be tested.

It is important to point out that the method currently used to test hypotheses is a hybrid of two approaches with radically different bases: one is the approach of Fisher and the other that of Neyman and Pearson (Cowles, 1989).

Further to the "null hypothesis significance test" proposed by Fisher (1925), Neyman and Pearson (1928a,b,

1933) developed the "hypothesis test" from a totally different perspective. These authors used probability theory to formalize a rule for deciding between two complementary hypotheses, the null hypothesis (H_0) and the alternative hypothesis (H_1), each of which had its own distribution. In this way, hypothesis testing took into account both $p(D/H_0)$ and $p(D/H_1)$. Moreover, Neyman and Pearson's approach distinguished between two types of error: Type I error (accepting H_1 when H_0 is true), with a probability α which must be fixed prior to carrying out the data analysis, and Type II error (accepting H_0 when H_1 is true), with probability β. This led to the introduction of new concepts such as power ($1–\beta$) and the critical region (decision rule). Generally speaking, while Fisher considered the null hypothesis significance test to be a method of inductive inference, Neyman and Pearson regarded the hypothesis test as a decision or behavior rule opposed to inductive inference. However, as Chow (1996) points out, there are shared metatheoretical assumptions underlying the two approaches; for example, both consider the test of statistical significance to be a scientific method (a detailed discussion of the differential characteristics of the two methods can be found in Chow, 1996, p. 21–23).

From a perspective which aims to integrate the approaches of Fisher and that of Neyman and Pearson, Chow (1996) argues that Fisher's conceptualization of the inferential process and Neyman and Pearson's decision statistic are both features of NHST, even though they refer to different stages of it. In his view NHST comprises: (a) a binary decision in which we choose between two possible ways of characterizing the data, that is, in which we decide whether the statistic has exceeded the critical value or not, this supplying the minor premise for; (b) a conditional syllogism through which we opt for an explanation based on the influence of chance (H_0) or the absence of chance (not-H_0) and whose conclusion, in the event that the H_0 is rejected, supplies the minor premise for; (c) a disjunctive syllogism through which we provisionally accept H_1. Thus, at the statistical level obtaining a statistically significant result with NHST implies that the calculated statistic is as or more extreme than the criterion value. At the conceptual level it means that taking into account the influence of chance as a plausible explanation for the data obtained, the probability of such data occurring is very low.

Main Criticisms of NHST

As was pointed out in the introduction several criticisms have been levelled against NHST, the following being particularly noteworthy.

1. *NHST does not provide the information which the researcher wants to obtain.* One of the strongest criticisms of NHST concerns the type of information it provides. Several authors (for example, Berger & Sellke, 1987; Carver, 1978; Cohen, 1990, 1994; Cronbach & Snow, 1977; Falk & Greenbaum, 1995; Gigerenzer & Murray, 1987; Kirk, 1996; Oakes, 1986; Rozeboom, 1960) have argued that NHST and statistical inference have different objectives. The aim of statistical inference is to know the probability that H_0 is true given the results or data obtained in the sample ($p(H_0/D)$). However, NHST only tells us about the probability of obtaining data which are equally or more discrepant than those actually obtained in the event that the H_0 is true ($p(D/H_0)$). In this regard, Lindley (1957) showed that under certain conditions the $p(H_0/D)$ could approach 1 while the $p(D/H_0)$ approached 0, thus illustrating that the p value does not reflect the probability that the H_0 is correct. This phenomenon, which has also been demonstrated by defenders of Bayesian statistics (for example, Edwards, 1965; Shafer, 1982), is known as Lindley's paradox.

Therefore, if probability is understood as the limit of relative frequency, obtaining a statistically significant result does not imply that the H_0 is improbable. This posterior probability can only be obtained through Bayesian statistics, on the basis of the probability of the H_0 prior to carrying out the study ($p(H_0)$) and by considering probability as the degree of belief in a given hypothesis. It should be noted that Nickerson (2000) identifies several misconceptions regarding *calculated p* and *critical p*, defined in the previous section, which are closely related to confusion between the two types of conditional probabilities highlighted by the present criticism.

2. *Logical problems derived from the probabilistic nature of NHST.* This criticism, which is shared by the authors mentioned in the previous criticism, among others, states that NHST is based on an incorrect application of syllogistic deductive reasoning, specifically, of the rule known as *modus tollens* (denying antecedents by denying consequents) because probabilistic statements are incompatible with the rules of deductive reasoning. This kind of faulty reasoning has been termed "the illusion of attaining improbability" (Falk & Greenbaum, 1995) or "the odds-against-chance fantasy" (Carver, 1978). Such reasoning is associated with several misconceptions, one of which was pointed out in the previous section, namely, that the p value is the probability that the H_0 is correct. The other false beliefs, which will be described in more detail below, are that the complementary value of *p, (1–p)*, expresses the probability that the H_1 is correct, and that statistically significant results will be obtained in the event that the experiment is replicated.

3. *NHST does not enable psychological theories to be tested*. Several authors (for example, Bakan, 1966; Carver, 1978; Cohen, 1987, 1990, 1994; Oakes, 1986) have pointed out that a misconception closely related to the previous criticism is that NHST can be used to determine the probability that a research hypothesis is correct, and consequently, that the theory behind it has been confirmed (Lykken, 1968; Cohen, 1994). According to Cohen (1994) and Rozeboom (1960) the dichotomous decision to reject or accept the H_0 does not enable a psychological theory to be tested. Similarly, Carver (1978), Erwin (1998), Nickerson (2000) and Snow (1998), among others, argue that even when a H_0 is rejected objectively, it is still necessary to exclude another series of alternative, competing hypotheses prior to verifying the validity of the research hypothesis. Thus, the increased truthfulness of this hypothesis can only come from a solid theoretical base, an appropriate research design and multiple replications of the study under different conditions.

Furthermore, many authors (for example, Bakan, 1996; Bracey, 1991; Cohen, 1994; Meehl, 1967; Nickerson, 2000; Rosenthal, 1983, 1993; Rosnow & Rosenthal, 1989; Shaver, 1985, 1993; Thompson, 1996; Thompson & Snyder, 1998; Wilson, Miller, & Lower, 1967) argue that NHST also fails to provide information about the practical importance of results and the magnitude of observed effects. As it does not provide quantitative information these authors believe that the approach is unable to identify the true relationship between the population parameters on the basis of the sample statistics, and also that it underestimates the importance of the magnitude both of the phenomena studied and the units in which these are measured. As Tukey has pointed out on more than one occasion (1969, 1991) the advance of knowledge requires information about both the direction of the difference as well as its magnitude; however, NHST only tells us about direction.

4. *The fallacy of replication*. Various authors (for example, Bakan, 1996; Carver, 1978, 1993; Cohen, 1994; Falk & Greenbaum, 1995; Gigerenzer, 1993; Gigerenzer & Murray, 1987; Lykken, 1968; Oakes, 1986; Shaver, 1993; Thompson, 1996) argue that another misconception linked to the illusion of attaining improbability is that the complementary value of *p*, (*1–p*), expresses the probability that the results are replicable. If *calculated p* enabled us to know what the probability was of H_0 being true in the population, then it would constitute an indicative index of the replicability of the results; however, as has already been pointed out *calculated p* does not provide this information. Indeed, in order to obtain an estimator of the probability of the sample statistics it is necessary to assume that the H_0 is exactly true in the population, as if we don't start from such an assumption regarding the population parameters there would be an infinite number of possible estimators of *p* and the answer to the question posed by the test of statistical significance would be mathematically indeterminate. Thus, NHST evaluates the probability of sample statistics assuming that the H_0 is exactly true with respect to the corresponding population parameters.

In fact, what Gigerenzer (1993) calls "the fallacy of replication" confuses the level of significance with statistical power. It is true that when the effect size and the sample size of a replica study coincide exactly with those of the original study and the H_0 is false, then there is a decreasing monotonic relationship between replicability and *p* values, as Greenwald et al. (1996) point out; however, this relationship is not maintained when the H_0 is true. Therefore, as we will see later, the authors who make this criticism consider that the most suitable method for determining whether a phenomenon is replicable or reliable is either to carry out an external replication of the original study or to apply strategies of internal replication such as jackknife or bootstrap procedures.

5. *NHST fails to provide useful information because H_0 is always false.* Another important criticism levelled at NHST (for example, by Bakan, 1966; Berkson, 1938; Binder, 1963; Cohen, 1990, 1994; Grant, 1962; Greenwald, 1993, Lindgren, 1976; Lykken, 1968; Meehl, 1967, 1990a,b; Murphy, 1990; Nunnally, 1960; Oakes, 1986; Pollard, 1993; Schmidt, 1992; Thompson, 1992; Tukey, 1991; Weitzman, 1984) is that the method is of no use as in the population the H_0 is always false. Consequently, the decision to reject it simply means that the research design is powerful enough to detect an effect which is known to exist, regardless of its magnitude or usefulness. Thus, obtaining a statistically significant result depends more on sample size than on the truth or falsity of the research hypothesis or the appropriateness of the theory on which it is based (Hays, 1994; Oakes, 1986). In this regard, Cohen (1962, 1990, 1994) and Kirk (1996), among others, consider it ironic that the ritual of applying NHST leads researchers to be concerned with controlling for Type I errors, which cannot actually occur given that all the H_0 are false, while at the same time allowing Type II errors to reach unacceptable levels, in the order of 0.50 to 0.80. Nickerson (2000) adds that this tendency is of no little importance as, in many applied contexts, the costs associated with a Type II error are greater than those resulting from a Type I error. Taking as their starting point the fact that the H_0 is always false, Cohen (1994) and Meehl (1967) conclude that use of NHST is only valid in true experiments which include randomization or when the slightest deviation from pure chance may be important.

6. *Problems associated with the dichotomous decision to reject/not reject the H_0*. NHST has also been criticized (for example, by Glass, McGaw, & Smith, 1981; Kirk, 1996; Rosnow & Rosenthal, 1989; Rozeboom, 1960) due to the fact that when researchers adopt a fixed significance level, they convert a continuum of uncertainty, which ranges from probability 0 to probability 1, into a dichotomous decision to reject/not reject the H_0. Moreover, the criterion used to choose the significance level which establishes the cut-off point around which results are either statistically significant or not is totally arbitrary (Gigerenzer, 1993; Glass, McGaw, & Smith, 1981; Johnson, 1999; Rossi, 1997).

7. *NHST impedes the advance of knowledge*. A final and general criticism of NHST (made, for example, by Cohen, 1987, 1994; Dar, 1987; Grant, 1962; Hunter, 1997; Schmidt, 1992, 1996; Schmidt & Hunter, 1997; Thompson, 1996; Tukey, 1991), and which follows directly from its inability to test theories (see criticism 3), is that the procedure impedes the advance and accumulation of theoretical knowledge. Starting from the notion that the H_0 is always false the above-mentioned authors argue that the fact that such a hypothesis is not rejected only means that the researcher is unable to specify the direction of the difference between certain conditions. In contrast, rejecting it indicates that the direction of this difference can be established with a certain degree of confidence. In the face of such an attitude, the critical authors consider that merely knowing the direction of a difference is not a sufficient basis for developing a psychological theory. In their view a project of such importance also requires that researchers establish the magnitude of the difference and the error associated with its estimate.

Alternatives to NHST

Various alternatives have been suggested with the aim of overcoming the problems associated with NHST; in the view of the more radical critics these should replace NHST, while more moderate voices argue that they should be a complement to significance testing. The present article focuses on those alternatives recommended by the TFSI of the APA (Wilkinson & TFSI, 1999) in the interests of improving the working methods of researchers with respect to statistical analysis and data interpretation; specifically, these are point estimates and confidence intervals, effect sizes, confidence intervals for the effect size, power analysis and replication.

Point Estimates and Confidence Intervals

The TFSI states that “it is hard to imagine a situation in which a dichotomous accept–reject decision is better than reporting an actual *p* value or, better still, a confidence interval” (p. 599).

A confidence interval, in accordance with a probability distribution, is used to test the confidence with which the true population value falls within a range of estimates. For example, a confidence interval of 95% indicates that if we repeatedly extract random samples from the population an indefinite number of times and then calculate a confidence interval for each sample, these intervals would include the estimated population parameter in 95% of the replications (Bleymüller, Gehlert, & Gülicher, 1988; Cohen, 1994). Therefore, with respect to sample data, confidence intervals are random variables, whose width and position varies from one replication to another.

Many authors (for example, Bakan, 1966; Brandstätter, 1999; Cohen, 1990, 1994; Hunter, 1997; Kirk, 1996, 2001; Loftus, 1991, 1995, 1996; Loftus & Masson, 1994; Meehl, 1997; Rozeboom, 1960; Schmidt, 1996; Schmidt & Hunter, 1997; Steiger & Fouladi, 1997) consider that calculating confidence intervals around estimates constitutes an excellent complement to, or even a substitute for, significance testing. These authors argue that such intervals provide information not only about the nil-null hypothesis, but also those H_0 which do not take the zero value (non-nil null hypothesis). Moreover, the confidence interval reflects the accuracy of the population parameter estimate. Thus, wide intervals give less accurate estimates than do narrow ones. In the case of confidence intervals for differences between parameters, these not only enable the hypothesis of no difference to be rejected when the interval does not include zero, but also indicate the direction and magnitude of the difference. Defenders of this alternative argue that it is as useful as NHST for deciding if chance or sample variability constitutes an improbable explanation of an observed difference. Moreover, point estimates and confidence intervals are governed by the same unit of measurement as the data, thus facilitating interpretation of the results. Two further noteworthy advantages of confidence intervals over NHST are that: (1) they enable the level of real error to be maintained at 0.05 (or at the level established in terms of a given confidence interval); and (2) they provide information that is highly useful for carrying out meta-analytic studies in the future.

In sum, confidence intervals avoid many of the problems inherent in classical significance tests. They do not require hypotheses to be formulated *a priori*, nor do they test trivial hypotheses. Furthermore, they provide more information and are easier to interpret than significance tests.

It is worth pointing out, with respect to this alternative, that Tryon (2001) proposes reformulating the way in which confidence intervals are traditionally constructed so that the new intervals enable inferences to be made about the existence of statistically significant differences, equivalence and indeterminacy between group means. In his opinion, this reformulation would avoid many of the problems derived from incorrect interpretations associated with NHST.

Effect Sizes

The TFSI urges researchers to present, in all cases, effect sizes for the main results. Moreover, special emphasis is placed on the need to interpret effect sizes within a practical and theoretical context, as well as the importance of such indices when carrying out analyses of power or meta-analyses in the future (p. 599).

Cohen (1988) defines the effect size as the extent to which the phenomenon is found within the population or, in the context of statistical significance testing, the degree to which the H_0 is false. For their part Snyder and Lawson (1993) argue that the effect size indicates the extent to which the dependent variable can be controlled, predicted and explained by the independent variable(s).

Many authors (for example, Brandstätter, 1999; Carver, 1978, 1993; Cohen, 1987; Cook & Campbell, 1979; Fisher, 1925; Folger, 1989; Glass, 1976; Harris, 1991; Kirk, 1996, 2001; Rosenthal, 1984; Rosnow & Rosenthal, 1989; Schmidt, 1996; Snyder & Lawson, 1993) consider that effect sizes should be calculated and interpreted in all research studies.

In addition to providing information about the magnitude of the observed effect, the effect size enables direct comparison of the results obtained in different studies, as these indices are transformations onto a common scale. They are also essential for carrying out analyses of power within the context of statistical significance tests, as well as for meta-analytic studies. Furthermore, when the measurement scales used for the variables are unfamiliar, standardized measures of effect size and their confidence intervals can provide information about the practical significance of the results obtained in a given study. However, it should be remembered that the use of standardized effect sizes is also subject to controversy (for example, Greenland, 1998).

Closely associated with the alternative based on the calculation of effect size, other authors (Cooper, 1979; Cooper & Rosenthal, 1980; Glass, 1976; Howard, Maxwell, & Fleming, 2000; Hunter & Schmidt, 1990; Schmidt, 1992, 1996) have suggested carrying out a meta-analysis whenever an attempt is made to analyse data extracted from multiple studies. In addition, and bearing in mind that it directly affects effect size, the TFSI stresses the importance of researchers assuring, or at least evaluating, the reliability of their data.

Confidence Intervals for the Effect Size

The TFSI recommends providing confidence intervals for all those effect sizes associated with main results. Moreover, it highlights the importance of researchers comparing confidence intervals across different studies rather than restricting themselves to verifying whether or not these intervals include the zero value. It also warns against making the common error of assuming that a parameter is included within a confidence interval (p. 599).

As has already been pointed out a confidence interval is an interval or range of possible population values which are reasonably consistent with the data observed in the sample. The level of confidence associated with the interval reflects the probability that, in the event that infinite confidence intervals are calculated for a given parameter, the intervals of the samples randomly taken from the population include the population value.

An increasing number of authors (for example, Cumming & Finch, 2001; Fidler & Thompson, 2001; Robinson & Wainer, 2001; Schmidt, 1996; Smithson, 2001; Thompson, 2002) are following the TFSI recommendations regarding calculation of confidence intervals for the effect size. However, the practice remains far from being a habitual one. This may be due to the fact that calculating confidence intervals for the effect size requires: (a) the use of noncentral distributions (Fisher, 1931; Pearson & Harley, 1972); and (b) specialized software able to estimate these intervals iteratively.

Those authors who recommend calculating confidence intervals for the effect size do so mainly according to the following arguments:

1) Confidence intervals for the effect size provide information that is readily understandable and which helps to interpret results appropriately.
2) There is a direct association between confidence intervals and NHST: when an interval excludes a given value it is necessary to reject the H_0 which states that this value is true, there being a certain level of significance related to the confidence level of the interval. This relationship between confidence intervals and the significance test may lead to a better understanding of the logic underlying both strategies.
3) Confidence intervals are highly useful for gathering empirical evidence across different studies. Thus, they promote the carrying out of meta-analyses.

4) The width of confidence intervals provides information about the accuracy of the estimate that is more useful and accessible than that gained through a value of statistical power.

Finally in this section, it should be noted that Rosenthal and Rubin (1994) proposed, as an alternative to confidence intervals, the calculation of counternull intervals for effect sizes. These authors define the counternull value of an effect size as the nonnull magnitude of the effect size which is exactly supported by the same amount of evidence as the null value of that size. In their opinion, the habitual calculation of the counternull value alongside the *p* value would eliminate the common error of considering that not rejecting the H_0 is equivalent to obtaining an effect size equal to zero. Moreover, it would help to eradicate the misconception that obtaining a statistically significant result is associated with achieving a scientifically important effect.

Power Analysis

As, according to the TFSI, power analysis is most meaningful when performed prior to gathering and examining data, they go on to recommend that a range of power analyses be carried out in order to observe how power estimates change with respect to different effect sizes and α levels. In addition, it is suggested that when describing results the calculated power should be replaced by confidence intervals (p. 596).

Power analysis arose out of the perspective on statistical decision-making proposed by Jerzy Neyman and Egon Pearson (Neyman & Pearson, 1928a, 1928b). These authors argued that given the magnitude of the difference between the null and alternative hypotheses (that is, the effect size in the hypothetical population), and by fixing values for the probabilities associated with Type I error (α) and Type II error (β), it was possible to determine the sample size necessary to detect an effect that actually existed in the population; alternatively, once the effect size, α and the sample size were fixed, it was possible to determine β, or its complement, the probability of rejecting the H_0 when it is false, namely, the power of the given test procedure. This analysis is considered particularly relevant when, it having proved impossible to reject the H_0, an attempt is made to conclude that there is either no effect or that its magnitude is insignificant (Meehl, 1991; Robinson & Levin, 1997; Schafer, 1993).

The findings of Cohen with respect to the limited power of most studies carried out in psychology (Cohen, 1962) continue to be relevant today (Kazdin & Bass, 1989; Rosnow & Rosenthal, 1989; Sedlmeier & Gigerenzer, 1989); in the opinion of critics of NHST, this constitutes a serious problem that hinders the advance of knowledge and which is closely linked to the wrong interpretations made regarding the information provided by significance testing.

Replication

The TFSI warns researchers that in their desire to reject the H_0 they may make the mistake of publishing false theories derived from the use of an inadequate methodology, even though the statistical analysis is correct. In addition, it is suggested that replications of the original study should be carried out in order to avoid this problem (p. 600).

As Allen and Preiss (1993) state, scientific knowledge is developed through replication. The results of an unreplicated study, regardless of the statistical significance achieved, are bound to be speculative (Hubbard & Armstrong, 1994) and lacking in any inherent meaning (Lindsay & Ehrenberg, 1993). However, as Hubbard and Ryan (2000) point out, although many authors (for example, Carver, 1978, 1993; Cohen, 1990, 1994; Falk & Greenbaum, 1995; Hubbard, 1995; Levin, 1998; Lykken, 1968; Robinson & Wainer, 2001; Rosnow & Rosenthal, 1989; Shaver, 1993; Thompson, 1993, 1994, 1996, 1997) highlight the fundamental role played by replication in the advance of knowledge, the percentage of articles concerning a replicated study which are published in psychology journals is scarce.

Those authors who defend replication argue that the attempt to reproduce the results of a previous study is an essential procedure for preventing the empirical literature from being plagued by spurious results, as the most objective method for checking whether the result of a single experiment is reliable is replication. Both external replication (carrying out new experiments) and internal replication (using methods such as crossed validation, or jackknife and bootstrap procedures) can serve to meet this objective.

In addition to the alternatives described in this section it should be pointed out that Thompson (1994, 1996) proposes modifying certain linguistic habits in order to avoid problems associated with a mistaken interpretation of the information provided by NHST. Specifically, he suggests using the expression *statistically significant* rather than the term *significant*, and argues that researchers should avoid saying things such as *the results reached statistical significance*. In agreement with this proposal, Nickerson (2000) also considers that certain aspects of the controversy surrounding NHST would cease to be meaningful if unambiguous expressions were used when describing outcomes.

Replies to the Criticisms of NHST

As pointed out in the Introduction, not all authors believe that NHST lacks usefulness in the field of research. Indeed, excellent papers have been published refuting each of the criticisms levelled against NHST. Moreover, the alternatives proposed to replace, or at least improve, the information provided by NHST are not exempt from criticism. This section discusses the arguments against each of the above-mentioned criticisms, as well as the shortcomings of some of the strategies proposed as possible alternatives to NHST.

1. *NHST does not provide the information which the researcher wants to obtain*. One of the most radical responses to this criticism is that of Hagen (1997), who takes as his starting point the article by Cohen (1994) which constitutes one of the strongest attacks against NHST. Hagen argues that in the example on which Cohen bases his critique an attempt is made to associate the probability of the H_0 with empirically-based and quantifiable relative frequencies, in such a way that the H_0 and the H_1 are considered to be statements about the sample. However, statistical hypotheses must always make reference to the population (Hayes, 1963). The H_0 proposed in Cohen's example does not provide any information about the sample distribution of the test statistic as it refers to the sample; it is therefore invalid as a H_0. Although Cohen states that the posterior probability of interest to the researcher can only be obtained through Bayesian statistics, Hagen argues that a method of probability based on relative frequencies cannot be used in a Bayesian analysis in order to establish the posterior probabilities of the H_0 and the H_1. In sum, Hagen's view is that if we start from the notion of probability based on relative frequencies (Fisher, 1935) the significance test does not provide the information the researcher wants, but if the probability of the H_0 is compared with the degree of subjective belief (Jeffrey, 1934) then the test does indeed provide the desired information. Other authors such as Baril and Cannon (1995), Cortina and Dunlap (1997) and McGraw (1995) also believe Cohen's example to be unsuitable, and demonstrate, moreover, that in certain situations the $p(D/H_0)$ and the $p(H_0/D)$ are not fundamentally different. In a similar vein, Nickerson (2000) argues that when the $p(H_1)$ is as high as the $p(H_0)$ and the $p(D/H_1)$ is much greater than the $p(D/H_0)$, a low value for the $p(D/H_0)$ enables the researcher to predict, to a high degree of accuracy, that the $p(H_0/D)$ will also have a low value. Cortina and Dunlap (1997) also maintain that this criticism does not reflect a shortcoming of NHST in itself, but rather a problem of interpretation with respect to the information it provides.

Focusing exclusively on this criticism, which he terms the converse inequality argument, Markus (2001) argues that even when NHST is not a formally valid procedure for testing a hypothesis its use can still be justified within an inductive framework.

Other authors, such as Frick (1996) and Chow (1996), accept the criticism but do not believe that Bayesian statistics are able to solve the underlying problem. Frick argues that the *a priori* calculation of probabilities is subjective and arbitrary and the response provided by Bayesian analysis is transitory. Chow is also severely critical of those authors who suggest that Bayesian statistics represent a solution to this problem. He argues that the Bayesian perspective on empirical research is highly questionable for two basic reasons: (a) it totally ignores whether an explicative hypothesis is consistent with the phenomenon it aims to explain; and (b) considering the Bayesian theorem as a mere inductive rule leads to empirical research being regarded as equivalent to pure formalism. According to Chow, the main objective of the empirical researcher should not be to calculate the degree of belief in the hypothesis prior to carrying out the study, nor to evaluate its consistency with the rest of the belief system, but rather to examine if the hypothesis is consistent with the phenomenon under study.

2. *Logical problems derived from the probabilistic nature of NHST*. Let us recall that this criticism is centered on the assumption that the rule of deductive reasoning known as *modus tollens* cannot be applied to probabilistic premises. NHST is based on the following sequence of premises: if the H_0 is true then a sample taken from the population associated with the null value will probably give a statistic located within a given range of values (that is, if A then probably B); the sample statistic does not fall within the range of values (that is, not B); consequently, the sample probably does not come from a population associated with the null value (that is, probably not A).

Unlike those authors who reject outright the validity of such reasoning, Cortina and Dunlap (1997) show that, under certain conditions, this sequence of premises does not violate the rules of syllogistic reasoning. Thus, although *modus tollens* cannot be applied to probabilistic premises when the truth value of the antecedent of the first premise is not related, or is negatively related, to the truth value of its consequent, it is a valid procedure in those cases, common within psychology, where the truth values of the two components, that is, of the antecedent and the consequent of the first premise, are positively related. From another perspective, Hagen (1997) maintains that a formally valid argument is not always adequate; indeed, an argument may be reasonable and sustainable even when it lacks logical validity in the formal sense. Thus, he questions whether logical validity is really an essential criterion for scientific argumentation.

With respect to the misconception that the *p* value is the probability that the results are due to chance, a notion closely linked to the criticism under discussion here, Chow (1996) argues that misinterpretation of the meaning of *p* would be easily rectified were researchers to understand that *p* is a conditional accumulated probability which depends on the H_0 being true.

3. *NHST does not enable psychological theories to be tested*. As Cortina and Dunlap (1997) point out, the aim of data analysis is to examine the extent to which the data are consistent with the theoretical responses to the questions posed in the research. This empirical testing must be objective and, in line with the Popperian approach (1959), able to refute all possible alternative hypotheses and explanations. In Cortina and Dunlap's view, when NHST is applied in the context of an appropriate experimental design it is one of the best analytic procedures for conducting this kind of testing.

Starting from a much less ambitious conception of the usefulness of NHST than that used by Cortina and Dunlap, Chow (1996) tackles this criticism by arguing that it is based on a clear confusion between statistical and nonstatistical questions. In line with the critics he believes that the research hypothesis is not directly comparable with the H_1; the result of significance testing in itself does not provide sufficient evidence to confirm the substantive hypothesis, it merely tells us whether or not there are rational grounds for excluding unknown random factors as plausible explanations of the data. However, it is necessary to determine which specific, nonrandom factor(s) is/are responsible for the results obtained, and this is not merely a statistical question but a matter of inductive inference. Thus, while it is true that NHST does not enable theories to be tested, neither do most of the alternatives proposed to replace it (for example, effect size or meta-analysis), because testing a theory involves much more than the refutation of a statistical hypothesis.

4. *The fallacy of replication.* In contrast to those authors who consider that the complementary value of *p*, (*1–p*), does not express the probability that the results are replicable, Greenwald et al. (1996) show that when the effect size and the sample size of the replication coincide exactly with those of the original study, the complementary value of *p* provides a measure of confidence in replicating the rejection of a H_0. However, it should be acknowledged that most authors recognize that the decreasing monotonic relationship between replicability and *p* values is not maintained when the H_0 is true.

Chow (1996) argues that in addressing the fallacy of replication identified by Gigerenzer (1993) it is important to take into account the mathematical basis of NHST, namely, the sample distribution of the statistic. Indeed, the nature of the sample distribution illustrates that there is nothing inherent within NHST that need encourage the wrong interpretation underlying the fallacy of replication.

5. *NHST fails to provide useful information because H_0 is always false*. It should be remembered that the main argument on which this criticism is based is that even when the different observed samples come from the same population they will always differ among themselves on any variable measured, and therefore in a literal sense the H_0 is always false.

In response to this criticism, Hagen (1997) argues that the H_0 does not propose equal samples, but rather, starting from the assumption that there are in fact differences, postulates that the samples have been taken from the same population. He adds that contrary to the opinion of those who make this criticism, when samples belong to the same population the probability of rejecting the H_0 does not approach 1 as the sample size increases.

For their part Cortina and Dunlap (1997) argue that use of the zero value associated with the H_0, even when the latter is false in a literal sense, is able to provide useful information. Indeed, following the good-enough principle of Serlin and Lapsley (1985, 1993), a strategy subsequently defended by Rouanet (1996) and Murphy and Myors (1999), the zero value can be taken as the mid-point of an interval which: (a) includes all the values which should be considered as trivial; and (b) is small enough to enable the calculations based on the zero value to provide a good estimate of the calculations based on other values belonging to the interval. From this perspective rejecting the H_0, in the context of significance testing, may indeed provide the researcher with relevant information.

Tackling the criticism head on, Baril and Cannon (1995) and Frick (1995) argue that the H_0 may be true. The latter author maintains that assigning a probability of occurrence other than zero to the zero value does not violate any rule of probability. He nevertheless distinguishes between those situations in which the H_0 may be true from those in which it cannot be. Thus, he considers that in purely applied experiments where complex variables are manipulated it is very difficult for the H_0 to be true. However, in experiments of a more theoretical nature where only one variable is manipulated the H_0 may be true. Obviously, in order to be able to accept the H_0, the results of the experiment must be consistent with it; thus, given that the statistic does not enable the truth value of the hypothesis to be tested Frick proposes that this objective be met by using the criterion of adequate effort. This criterion means that researchers apply all those methodological strategies which increase the probability of detecting an effect which actually exists.

In Chow's (1996) opinion it is not assumed in the present criticism that the H_0 forms part of a conditional proposition, but rather that it is a categorical proposition, and this undermines the validity of the criticism. Indeed, in NHST the H_0 is used on two occasions as a component of a conditional proposition. Thus, the H_0 constitutes the necessary condition for accepting the random variations as a plausible explanation of the data in the following proposition: *if chance explains the results, then H_0.* Moreover, it is put forward as the sufficient condition for proposing a given sample distribution: *if H_0, then the statistic is distributed according to a sample distribution of the difference whose difference of means is zero*. Furthermore, the fact that both Meehl (1967) and Cohen (1994) restrict their critique to the nonexperimental field illustrates that their arguments are not merely statistical, but also address aspects associated with research design.

6. *Problems associated with the dichotomous decision to reject/not reject the H_0.* In contrast to those authors who criticize the arbitrariness involved in choosing the α value that determines the cut-off point used to decide whether results are statistically significant or not, Cox (1977) and Frick (1996) maintain that this criterion has been adequately established by the scientific community; furthermore, it enables researchers to eliminate the influence of their judgments and opinions when interpreting data and thus guarantees objectivity. In a similar vein, Chow (1996) argues that it is an objective criterion at a mathematical level, one whose meaning is not linked to the theoretical background of the researcher. It constitutes an unambiguous index which reflects the rigor adopted by the researcher in deciding to reject chance as an explanation of the data. The criticism of Gigerenzer (1993) that, given its mechanical and conventional nature, NHST has become institutionalized, a statistically significant result now being regarded as indicative that a study has been properly carried out, is not, in Chow's (1996) view, a good reason for researchers to cease rejecting or not rejecting the H_0 in accordance with a criterion based on an arbitrary value. However, it should make us aware of the need to use this decision-making criterion within a suitably valid research design.

Furthermore, although the choice of the α value is widely considered within the scientific community to be arbitrary, it should be pointed out that in their analysis of the history of statistical theory and probability Cowles and Davis (1982) argue that this choice was not in fact arbitrary, but rather was derived from scientific conventions based on the notion of chance and the improbability of the occurrence of a given event (Pearson, 1900; Student, 1908).

7. *NHST impedes the advance of knowledge*. One of the authors who has most strongly argued that NHST favors the advance of knowledge is Frick (1996). In his opinion, it is the optimum method for obtaining sufficient empirical evidence in support of what he calls "ordinal statements." He defines such statements as those which only specify the order of the conditions or the effects, or the direction of a correlation, although he adds that the statistical operations used to justify these statements usually assume a scale that is more precise than a mere ordinal one. As, according to Frick, most theories and laws tested in psychology are ordinal, NHST is a suitable method for distinguishing between those findings which should form part of standard psychological knowledge and those which are not valid enough to enter this body of theoretical knowledge. He also argues that in experiments whose aim is the immediate application of results, effect size is important and, therefore, the conclusions drawn from such experiments cannot be based solely on NHST. However, he believes that NHST is also valid in applied experiments. He acknowledges, as Berkson (1938) and Grant (1962) point out, that NHST is not valid for testing theoretical models in which quantitative predictions are made. In agreement with Frick, Abelson (1997) argues that NHST is particularly useful when the aim is to clarify whether a given difference between conditions is positive or negative.

In a good example of his deeply reflective style, Chow (1996) argues that those authors who make this criticism assume that: (1) quantitative information is more important than qualitative information; and (2) as effect size increases, more evidence is obtained in favor of the hypothesis being tested. In his opinion these assumptions are acceptable provided that: (1) the experimental hypothesis paraphrases the substantive hypothesis; (2) the aim of the experiment is to apply immediately the results obtained; and (3) the actions to be taken on the basis of the study results depend directly on the effect size. However, the criticism is not applicable to those experiments which aim to test a theory, given that in such cases the above-mentioned assumptions are not valid. In sum, the author shows, as he has done on other occasions, that the criticism goes beyond the statistical field addressed by NHST.

Generally speaking, most defenders of NHST believe it has been misinterpreted and badly used for decades. With this as his starting point, Hagen (1997) argues that the logic of NHST is elegant and creative and is perfectly integrated within the process of statistical inference. Krueger (2001) points out that even if we acknowledge that many of the criticisms regarding the lack of logical validity of inferences derived from NHST are pertinent, such inferences have an undeniable practical validity. Similarly, both Abelson (1997) and Dixon (1998) consider that NHST is able to provide information that enables us to answer important questions in the research field. From a different perspective, Chow (1996, 1998a,b) highlights the fact that most criticisms of NHST refer to nonstatistical problems which are derived from an incorrect interpretation of data gathered through nonexperimental methods. Moreover, he argues that some of these criticisms go beyond the field of research methodology; therefore he maintains that decision of whether such criticisms are pertinent or not should take into account the degree to which they are limited to the field of statistical conclusion validity, this being the field of influence of NHST.

Criticisms of the Alternatives Proposed to Replace or Improve the Information Provided by NHST

Various authors (for example, Abelson, 1997; Cortina & Dunlap, 1997; Hagen, 1997; Hayes, 1998) argue that confidence intervals do not solve the problems associated with NHST as they are based on the same logic. Indeed, rather than starting from a hypothetical parameter and establishing a sample distribution with which to compare the sample statistic, this approach establishes a confidence interval and compares an infinite number of parameters, taking the interval as a reference point. Cortina and Dunlap (1997) add that the calculation of confidence intervals is as imperfect a procedure as NHST. Thus, when a confidence interval is established there is also probability α of committing an error. Obtaining a confidence interval of 100% would only be possible were α equal to zero, in which case the range would be from $-\infty$ to $+\infty$ (or from -1 to $+1$ for the correlations), and such an interval would be of no use. Frick (1996) argues that when NHST is used to test an ordinal statement, the *p* value associated with the H_0 contains very important information and should therefore not be replaced by a confidence interval. Finally, Chow (1996) points out that although the critics argue that confidence intervals provide much more information than NHST, they do not explain how establishing such intervals helps to test theories in a different way to NHST.

With respect to the alternative based on the calculation of effect sizes, several authors argue that as the estimation of effect size depends on the variability of the measures and the experimental manipulations used with a given sample, it should be interpreted with caution (Brandstätter, 1999; Cortina & Dunlap, 1997; Dooling & Danks, 1975). The calculation of standardized effect sizes has also been widely criticized because, under certain circumstances, the standardization may distort the order or the intensity of the observed effects (Greenland, 1998). Moreover, the choice of the most suitable index for calculating effect size in a given context is a complex matter and authors tend to disagree on how it should be done (Crow, 1991; Gorsuch, 1991; McGraw, 1991; Parker, 1995; Strahan, 1991). Chow (1996) argues that behind the proposal to evaluate research results according to effect sizes lies the controversy between two types of evaluation criterion: the statistical criterion aimed at evaluating the influence of chance, and a series of nonstatistical criteria whose aim is to evaluate the impact of the results in real life. Defenders of NHST do not object to the need to use nonstatistical criteria in order to analyse the results in more detail, once the effects of chance have been ruled out; however, the defenders of the alternative based on the calculation of effect sizes are willing to use nonstatistical criteria even when it is not possible to exclude chance as a plausible explanation of the results. Furthermore, decisions made on the basis of effect sizes are as arbitrary and conventional as those based on the *p* value. With respect to another issue related to the alternative under discussion here, Chow (1996) also criticizes *meta-analysis* for considering that the accumulation of data, which forms the basis of this strategy, does not in itself favor the development of knowledge.

Finally, it should be pointed out that although Robinson and Wainer (2001) recommend calculating the effect size as a complement to NHST, they add that under certain circumstances obtaining such an index can prove very difficult, may be of no help in interpreting the data, and can even lead to wrong interpretations regarding the importance and/or accuracy of the results. These authors therefore argue that it should not be calculated in such cases.

In terms of the alternative based on power analysis, Hagen (1997) argues that it follows the same logic as that underlying NHST. Chow (1996) adds that the power value deemed adequate within the scientific community has also been established in an arbitrary and conventional way. Moreover, he believes that there are good reasons to question the validity of statistical power itself. Firstly, when power is taken into account the meaning of Type II error is modified. Secondly, it is not possible to represent graphically statistical power without wrongly representing the significance test. Thus, in Chow's opinion the probability of committing a Type II error (β) is a conditional probability which should be defined as *p(accept chance/not-H_0)*. However, in the framework of power analysis it is defined as *p(accept chance/H_1)*, which is wrong given that the H_1 cannot be considered to be equivalent to a *not-H_0*. Thus, while the H_1 plays an essential role in terms of how β is defined in the power analysis, it is not taken into account when defining Type II error in the context of statistical significance testing. Another relevant issue is that in a power analysis, power is the complement of a conditional probability which assumes that the H_1 is true prior to knowing the value of the Type II error, and thus the concept of power is based on a conceptual error. Indeed, the definition of the test's power as the probability that the H_1 is true is incorrect if it is necessary to assume *a priori* that this hypothesis is true.

Discussion

The present article has focused on the current controversy surrounding NHST. The most important criticisms of NHST, the main alternatives proposed to replace or improve it, and the arguments in defence of NSHT procedure's validity have all been described. On the basis of our analysis of the controversy we believe that three measures need to be adopted in order to increase the rigor and seriousness of research activity: (1) use NHST in the appropriate context; (2) complement the use of such testing with procedures which provide information about those aspects beyond the scope of NHST; and (3) follow a series of recommendations in order to foster the rational use of applied methodology in psychological research. Each of these measures is discussed below.

The first measure results from the conclusion that many of the criticisms levelled against NHST do not highlight shortcomings of the procedure itself, but rather derive from its incorrect use by researchers; this is mainly due to misconceptions regarding the type of information it provides. Thus, it seems to us that there is an urgent need to promote measures which enable researchers to understand that NHST only provides information about whether there is a rational basis for excluding sample error as a plausible explanation for the data, this being a step prior to the search for specific nonrandom factors able to explain the results. In sum, the use and scope of NHST should be restricted purely to the statistical field.

However, this does not prevent NHST being complemented by other procedures which enable other kinds of information about the data to be obtained. Thus, without wishing to reiterate what has already been stated elsewhere, we agree with those authors who recommend use of point estimates and confidence intervals as, unlike NHST, these intervals do provide information regarding the accuracy of the estimate of population parameters. In addition, we believe that provided researchers start from a detailed theoretical knowledge of the object of study, the calculation of effect sizes and their confidence intervals can be very useful in obtaining information about the practical value of the research results. Contextualizing these indices within a wider range of studies is also an excellent strategy for increasing the accuracy of estimates of the population parameters. With respect to power analysis we believe, unlike Chow (1996), that it is a very useful procedure for ensuring confidence in the results obtained. The same can be said for replication, although given the effort required (particularly in the case of external replication) we doubt that this procedure will come to be widely used within the scientific community.

The third measure we suggest adopting concerns a set of recommendations which we believe are important in terms of promoting the rational use of methodology in the field of psychological research. The first recommendation is that all studies are planned in great detail, and that researchers use a design which offers a good level of validity. Among other aspects, this means choosing a suitable sample size, controlling for possible sources of extraneous variation, manipulating or selecting appropriately the independent or predictive variables, and using measures which are sensitive enough to detect those effects which actually exist. Secondly, it is important to carry out a detailed graphic analysis of the data in order to evaluate whether or not they are consistent with the assumptions of the statistical model (Cleveland, 1993; Cleveland & McGill, 1988; Tufte, 1983, 1990; Tukey, 1962, 1977; Wainer & Thissen, 1993; Wilkinson & TFSI, 1999). The next step should then be the analysis of possible relationships between the variables, applying in a reflective way the statistical procedures described in this paper. This is why it is important to know what information is provided by each of these methods, the contexts and circumstances in which they can feasibly be applied, and the type of relationships that exist between them. At this point, it should not been forgotten the importance of providing graphics including essential information to get a better understanding of the set of data in hand (Cohen, 1994; Loftus, 1993; Tukey, 1962, 1977; Wilkinson & TFSI, 1999). In line with the recommendation based on the reflective use of statistical procedures it is important to highlight, as do the TFSI and various authors (for example, Cohen, 1990; Cortina & Dunlap, 1997; Falk & Greenbaum, 1995; Gigerenzer, 1993; Greenwald et al., 1996; Haller & Krauss, 2002), that such procedures should never be regarded as a substitute for the common sense and good judgment of the researcher.

Finally, and in accordance with what some authors (for example, Fidler, 2002; Hubbard & Ryan, 2000; Kirk, 1996; Robinson & Wainer, 2001; Schmidt, 1996) have already proposed in order to eradicate the dogmatic use of NHST, it only remains to point out that putting into practice the measures proposed here will only be possible through a multifaceted approach involving the authors of text books, university lecturers responsible for teaching undergraduate and postgraduate research methodology, the authors of doctoral theses, creators of statistical software, the editors and reviewers of scientific journals, and those bodies responsible for producing manuals of scientific publication guidelines, such as the *Publication Manual of the American Psychological Association* (APA).

Acknowledgments

This study was supported by funds from the Spanish Department of Science and Technology (Grant code: BSO2001–3751 C02–02).

References

Allen, M., & Preiss, R. (1993). Replication and meta-analysis: A necessary connection. *Journal of Social Behavior and Personality, 8*(6), 9–20.

Abelson, R.P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

Abelson, R.P. (1997). A retrospective on the significance test ban of 1999 (if there were no significance tests, they would be invented). In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117–144). Hillsdale, NJ: Erlbaum.

American Psychological Association (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

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

Baril, G.L., & Cannon, J.T. (1995). What is the probability that null hypothesis testing is meaningless? *American Psychologist, 50,* 1098–1099.

Berger, J.O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *P* values and evidence. *Journal of the American Statistical Association, 82*, 112–122.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the χ^2 test. *Journal of the American Statistical Association, 33*, 526–542.

Binder, A. (1963). Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 70*, 107–115.

Bleymüller, J., Gehlert, G., & Gülicher, H. (1988). *Statistik für Wirtschaftswissenschaften (5. Aufl).* München: Vahlen.

Bracey, G.W. (1991). Sense, non-sense, and statistics. *PhiDelta Kappan, 73*, 335.

Branstätter, E. (1999). Confidence intervals as an alternative to significance testing. *Methods of Psychological Research Online, 4*(2), 33–46.

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

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

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

Chow, S.L. (1987). *Experimental psychology: Rationale, procedures and issues*. Calgary, Alberta, Canada: Detselig Enterprises.

Chow, S.L. (1988). Significance test or effect size? *Psychological Bulletin, 103*, 105–110.

Chow, S.L. (1989). Significance tests and deduction: Reply to Folger (1989). *Psychological Bulletin, 106*, 161–165.

Chow, S.L. (1991). Some reservations about power analysis. *American Psychologist, 46*, 1088–1089.

Chow, S.L. (1996). *Statistical significance: Rationale, validity, and utility*. Beverly Hills, CA: Sage.

Chow, S.L. (1998a). Précis of statistical significance: Rationale, validity, and utility. *Behavioral and Brain Sciences, 21*, 169–239.

Chow, S.L. (1998b). What statistical significance means. *Theory and Psychology, 8*, 323–330.

Cleveland, W.S. (1993). *Visualizing data*. Summit, NJ: Hobart.

Cleveland, W.S., & McGill, M.E. (Eds.) (1988). *Dynamic graphics for statistics*. Belmont, CA: Wadsworth.

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

Cohen, J. (1987). *Statistical power analysis for the behavioral sciences* (rev. ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cooper, H.M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology, 37*, 131–146.

Cooper, H.M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin, 87*, 442–449.

Cortina, J.M., & Dunlap, W.P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2*, 161–172.

Cowles, M. (1989). *Statistics in psychology: An historical perspective*. Hillsdale, NJ: Erlbaum.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist, 37*, 553–558.

Cox, D.R. (1977). The role of significance tests. *Scandinavian Journal of Statistics, 4*, 49–70.

Cronbach, L.J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 116–127.

Cronbach, L.J., & Snow, R.E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

Crow, E.L. (1991). Response to Rosenthal's comment "How are we doing in soft psychology." *American Psychologist, 46*, 1083.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532–574.

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. *American Psychologist, 42*, 145–151.

Dar, R., Serlin, R.C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62*, 75–82.

Dixon, P. (1998). Why scientists value *p* values. *Psychonomic Bulletin and Review, 5*, 390–396.

Dooling, D., & Danks, J.H. (1975). Going beyond tests of significance: Is psychology ready? *Bulletin of the Psychonomic Society, 5*, 15–17.

Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin, 63*, 400–402.

Erwin, E. (1998). The logic of null hypothesis testing. *Behavioral and Brain Sciences, 21*, 197–198.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning, 9,* 83–96.

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

Fidler, F. (2002). The fifth edition of the APA publication manual: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement, 62*, 749–770.

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed-and random-effects effect sizes. *Educational and Psychological Measurement, 61*, 575–604.

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. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.

Fisher, R.A. (1931). Introduction. In J.R. Airey (Ed.), *Table of Hh functions* (pp. xxvi–xxxv). London: British Association.

Fisher, R.A. (1935). *The design of experiments*. London: Oliver & Boyd.

Folger, R. (1989). Significance tests and the duplicity of binary decisions. *Psychological Bulletin, 106*, 155–160.

Frick, R.W. (1995). Accepting the null hypothesis. *Memory & Cognition*, *23*(1), 132–138.

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

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 science: Volume 1. Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum.

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

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

Glass, G.V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher, 5*, 3–8.

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

Gorsuch, R.L. (1991). Things learned from another perspective (so far). *American Psychologist, 46*, 1089–1090.

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

Greenland, S. (1998). Meta-analysis. In K. Rothman & S. Greenland (Eds.). *Modern epidemiology*. Philadelphia: Lippincott-Raven.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20.

Greenwald, A.G. (1993). Consequences of prejudice against the null hypothesis. In G. Kerens, & C. Lewis (Eds.). *A handbook for data analysis in the behavioral sciences: Volume 1. Methodological issues* (pp. 419–448). Hillsdale, NJ: Erlbaum.

Greenwald, A.G., Gonzalez, R., Harris, R.J., & Guthrie, D. (1996). Effect sizes and *p*-values: What should be reported and what should be replicated? *Psychophysiology, 33*, 175–183.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis, 1*, 3–10.

Hagen, R.L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52*(1), 15–24.

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

Harris, R.J. (1991). Significance tests are not enough: The role of effect-size estimation in theory corroboration. *Theory and Psychology, 1*, 375–382.

Hayes, W.L. (1963). *Statistics for psychologists.* New York: Holt, Rinehart & Winston.

Hayes, A.F. (1998). Reconnecting data analysis and research designs: Who needs a confidence interval? *Behavioral and Brain Sciences, 21*, 203–204.

Hays, W.L. (1994). *Statistics* (4th ed.). New York: Holt, Rinehart and Winston.

Howard, G.S., Maxwell, S.E., & Fleming, K.J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and bayesian analysis. *Psychological Methods, 5*, 315–332.

Hubbard, R. (1995). The Earth is highly significantly round (p < .0001). *American Psychologist, 50*, 1098.

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

Hubbard, R., Parsa, A.R., & Luthy, M.R. (1997). The spread of statistical significance testing in psychology: The case of the *Journal of Applied Psychology*, 1917–1994. *Theory and Psychology, 7*, 545–554.

Hubbard, R., & Ryan, P.A. (2000). The historical growth of statistical significance testing in psychology and its future prospects. *Educational and Psychological Measurement, 60*, 661–681.

Huberty, C.J. (1987). On statistical testing. *Educational Researcher, 16*(8), 4–9.

Hunter, J.E. (1997). Need: A ban on the significance test. *Psychological Science, 8*, 3–7.

Hunter, J.E., & Schmidt, F.L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Jeffreys, H. (1934). Probability and scientific method. *Proceedings of the Royal Society of London, Series A, 146*, 9–16.

Johnson, D.H. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management, 63,* 763–772.

Kazdin, A.E., & Bass, D.(1989). Power to detect differences between alternative treatments in comparative psychotherapy outcome research. *Journal of Consulting and Clinical Psychology, 57*, 138–147.

Kirk, R.E. (1996). Practical significance: a concept whose time has come. *Educational and Psychological Measurement, 56*, 746–759.

Kirk, R.E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61*, 213–218.

Krueger, J. (2001). Null hypothesis significance testing. On the survival of a flawed method. *American Psychologist, 56*, 16–26.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist, 43*, 635–642.

Levin, J.R. (1998). To test or not to test H$_0$? *Educational and Psychological Measurement, 58*, 313–333.

Lindgren, B.W. (1976). *Statistical theory* (3rd ed.). New York: Macmillan.

Lindley, D.V. (1957). A statistical paradox. *Biometrika, 44*, 187–192.

Lindsay, R.M., & Ehrenberg, A.S.C. (1993). The design of replicated studies. *American Statistician, 47*, 217–228.

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

Loftus, G.R. (1993). A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments and Computers, 25*, 250–256.

Loftus, G.R. (1995). Data analysis as insight: Reply to Morrison and Weaver. *Behavior Research Methods, Instruments and Computers, 27*, 57–59.

Loftus, G.R. (1996). Psychology will be a much better science when we change the way to analyse data. *Current Directions in Psychological Science, 5*, 161–171.

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

Lykken, D. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*, 151–159.

Markus, K.A. (2001). The converse inequality argument against tests of statistical significance. *Psychological Methods, 6*, 147–160.

McGaw, K.O. (1991). Problems with the BESD: A comment on Rosenthal's "How are we doing in soft psychology?" *American Psychologist, 46*(10), 1084–1086.

McGaw, K.O. (1995). Determining false alarm rates in null hypothesis testing research. *American Psychologist, 50*, 1099–1100.

Meehl, P.E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103–115.

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

Meehl, P.E. (1990a). Appraising and amending theories: The strategy of Lakatosian defence and two principles that warrant it. *Psychological Inquiry, 1*, 108–141.

Meehl, P.E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195–244.

Meehl, P.E. (1991). Why summaries of research on psychological theories are often uninterpretable. In R.E. Snow, & D.E. Wilet (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach.* (pp. 13–59). Hillsdale, NJ: Erlbaum.

Meehl, P.E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 391–423). Hillsdale, NJ: Erlbaum.

Morrison, D.E., & Henkel, R.E. (1970) (Eds.), *The significance test controversy: A reader*. Chicago: Aldire.

Murphy, K.R. (1990). If the null hypothesis is impossible, why test it? *American Psychologist, 45*, 403–404.

Murphy, K.R., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology, 84*, 234–248.

Neyman, J., & Pearson, E.S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20A*, 175–263.

Neyman, J. & Pearson, E.S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika, 20A*, 264–294.

Neyman, J., & Pearson, E.S. (1933). On the testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society, 28*, 492.

Nickerson, R.S. (2000). Null hypothesis significance testing: A review of and old and continuing controversy. *Psychological Methods, 5,* 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 social and behavioral sciences*. New York: Wiley.

Parker, S. (1995). The "difference of means" may not be the "effect size." *American Psychologist, 50*, 1101–1102.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, 50*, 157–175.

Pearson, E., & Hartley, H. (1972). *Biometrika tables for statisticians* (Vol. 2). Cambridge, UK: Cambridge University Press.

Pollard, P. (1993). How significant is "significance"? In G. Keren, & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Volume 1. Methodological issues*. Hillsdale, NJ: Erlbaum.

Popper, K.R. (1959). *The logic of scientific discovery*. New York: Basic Books.

Robinson, D., & Levin, J. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26*(5), 21–26.

Robinson, D.H., & Wainer, H. (2001). *On the past and future of null hypothesis significance testing*. Princeton: Statistics & Research Division.

Rosenthal, R. (1983). Assessing the statistical and social importance of the effects of psychotherapy. *Journal of Consulting and Clinical Psychology, 51*, 4–13.

Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.

Rosenthal, R. (1993). Cumulating evidence. In G. Keren, & C. Lewis (Eds.), *A handbook of data analysis in the behavioral sciences: Volume 1. Methodological issues* (pp. 519–559). Hillsdale, NJ: Erlbaum.

Rosenthal, R., & Rubin, D.B. (1994). The counternull value of an effect size: A new Statistic. *Psychological Science, 5*, 329–334.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276–1284.

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.

Rossi, J.S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What* *if there were no significance tests?* (pp. 175–197). Hillsdale, NJ: Erlbaum.

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

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

Schmidt, F.L. (1992). What do data really mean? *American Psychologist, 47*, 1173–1181.

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

Schmidt, F.L. (2002). Are there benefits from NHST? *American Psychologist, 57*, 65–71.

Schmidt, F.L., & Hunter, J.E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Hillsdale, NJ: Erlbaum.

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

Serlin, R.C., & Lapsley, D.K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40*, 73–83.

Serlin, R.C., & Lapsley, D.K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren, & C. Lewis (Eds.), *A handbook of data analysis in behavioral sciences: Volume 1. Methodological issues*, (pp. 199–228). Hillsdale, NJ: Erlbaum.

Schafer, W.D. (1993). Interpreting statistical significance and nonsignificance. *Journal of Experimental Education, 61*, 383–387.

Shafer, G. (1982). Lindley's paradox. *Journal of the American Statistical Association, 77*, 325–334.

Shaver, J. (1985). Chance and nonsense: A conversation about interpreting tests of statistical significance. *PhiDelta Kappan, 67(1)*, 138–141.

Shaver, J. (1993). What statistical significance testing is, and what is not. *Journal of Experimental Education, 61*, 293–316.

Snyder, P. & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61*, 334–349.

Snow, R.E. (1998). Inductive strategy and statistical tactics. *Behavioral and Brain Sciences, 21*, 219.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61*, 305–632.

Steiger, J.H., & Fouladi, R.T. (1997). Noncentrally interval estimation and the evaluation of statistical models. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–258). Hillsdale, NJ: Erlbaum.

Strahan, R.F. (1991). Remarks on the binomial effect size display. *American Psychologist, 46*, 1083–1084.

Student [W.S. Gosset] (1908). The probable error of a mean. *Biometrika, 6*, 1–25.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Consulting and Clinical Psychology, 70*, 434–438.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education, 61*, 361–377.

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

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

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher, 26*(5), 29–32.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80*, 64–71.

Thompson, B., & Snyder, P.A. (1998). Statistical significance and reliability analyses in recent Journal of Counseling & Development research articles. *Journal of Counseling and Development, 76*, 436–441.

Tryon, W.W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*, 371–386.

Tufte, E.R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Tufte, E.R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

Tukey, J.W. (1962).The future of data analysis. *Annals of Mathematical Statistics, 33*, 1–67.

Tukey, J.W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*, 83–91.

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

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

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

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103–118.

Weitzman, R.A. (1984). Seven treacherous pitfalls of statistics, illustrated. *Psychological Reports, 54*, 355–363.

Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604.

Wilson, W., Miller, H.L., & Lower, J.S. (1967). Much ado about the null hypothesis. *Psychological Bulletin, 68*, 188–196.

Address for correspondence

Nekane Balluerka
Dpto. de Psicología Social y Metodología
de las Ciencias del Comportamiento
Facultad de Psicología
Universidad del País Vasco
Avda. de Tolosa, 70
E-20018 San Sebastián
Spain
Tel. +34 943018339
E-mail pspbalan@sc.ehu.es