

# Calculating the main alternatives to null-hypothesis-significance testing in between-subject experimental designs

Nekane Balluerka Lasa, Ana Isabel Vergara Iraeta and Jaume Arnau Gras\*

Universidad del País Vasco and \* Universidad de Barcelona

The present article deals with the controversy about null hypothesis significance testing (NHST) that currently exists in psychological research. The main criticisms and counter-criticisms posed by its detractors and supporters are presented. Alternatives have been proposed to replace or complement the NHST, as recommended by the Task Force on Statistical Inference of the APA, and these are explained. Furthermore, the procedure for calculating these alternatives using SPSS statistical software in a two-way between-factor design is demonstrated. Thus, an attempt is made to provide the applied researcher with resources that make it possible to analyse and interpret the results of any research study using a group of indicators that lends a high level of validity to the statistical inference performed.

*Cálculo de las principales alternativas a la prueba de significación de la hipótesis nula en diseños experimentales intersujetos.* El presente artículo se centra en la controversia existente actualmente en torno a la prueba de significación de la hipótesis nula en la investigación psicológica. Se exponen las principales críticas y contra-críticas que plantean sus detractores y defensores, así como las alternativas que, contando con el apoyo de la Task Force on Statistical Inference de la APA, han sido propuestas para sustituirla o complementarla. Además, se muestra la forma en la que pueden calcularse, mediante el programa estadístico SPSS, tales índices alternativos en un diseño factorial aleatorio. Con ello se pretende dotar al investigador aplicado de una serie de recursos que le permitan analizar e interpretar los resultados de cualquier estudio manejando un conjunto de indicadores que proporcionan un alto grado de validez a la inferencia realizada.

The inference revolution that took place in psychology between 1940 and 1955 led NHST (*Null Hypothesis Significance Testing*) to become the inductive inference method par excellence. This fact is made obvious by the strong increase in the percentage of *p* values used in empirical research since the 1950's, when 70% of the studies already based their analyses on the aforementioned index, reaching a percentage higher than 90% from 1990 on (Hubbard, Parsa, & Luthy, 1997; Hubbard & Ryan, 2000).

However, NHST has not been, nor is it currently, exempt from controversy. Several authors have highlighted the problems associated with the use of this method throughout various decades (for example, Bakan, 1966; Carver, 1978; Cohen, 1990, 1994; Finch, Cumming, & Thomason, 2001; Grant, 1962; Huberty, 1987; Kirk, 1996, 2001; Loftus, 1996; Meehl, 1990a, 1990b; Oakes, 1986; Rozeboom, 1960; Schmidt, 1996; Thompson, 1996), although excellent studies defending the validity and usefulness of this significance test have also been published (for example, Abelson, 1995; Chow, 1987, 1988, 1989, 1991, 1996, 1998a, 1998b; Cortina & Dunlap, 1997; Cox, 1997; Dixon, 1998; Frick, 1996; Hagen, 1997; Mayo, 1996).

In any case, although the debate surrounding the significance test generates a great deal of interest among scholars working on statistical and methodological issues, users of statistics applied to different areas of psychology are not aware of this controversy. In fact, a review of some of the recent literature shows that the exclusive use of the significance test still characterizes the majority of the studies published in psychology and that the development of alternative indices is occurring much more slowly than would be expected if the debate had a real effect. Besides other reasons pointed out in the conclusions of the present article, we consider that one of the main reasons for this inertia is the fact that researchers do not have guidelines to help them calculate alternative indices to the significance test with the statistical software they currently use.

Taking this context as a starting point, two aims are pursued in the present work. First, we try to return to the existing controversy surrounding NHST, so as to describe the current situation of the debate in our discipline. In order to do so, we will review the main criticisms put forward regarding the significance test, the arguments refuting each of these criticisms, and the alternatives, either complementary or replacement, whose use is recommended by the *Task Force on Statistical Inference (TFSI)* of the APA (Wilkinson & TFSI, 1999) to improve the working method of researchers within the framework of statistical analysis and data interpretation. Second, we try to lay the foundations for broadening the debate beyond the methodological field, providing the applied researcher with strategies that make it possible to calculate the main alternative indices to the significance test with

---

Fecha recepción: 21-2-08 • Fecha aceptación: 10-9-08

Correspondencia: Nekane Balluerka Lasa

Facultad de Psicología

Universidad del País Vasco

20080 Donostia (Spain)

E-mail: nekane.balluerka@ehu.es

statistical software widely used for psychological research, namely the SPSS. With this goal in mind, the data from a random factorial design are taken as reference, and the estimates of the main indices that make it possible to complement, or even replace, the information provided by the significance test are illustrated.

#### Main arguments for and against the NHST

In the present section, we will put forward the main arguments posed by the detractors and supporters of the significance test (for an exhaustive review of the existing controversy surrounding the significance test, see Balluerka, Gómez, & Hidalgo, 2005).

#### *Type of information provided*

One of the strongest criticisms of this test focuses on the type of information provided. Thus, several authors (for example, Carver, 1978; Cohen, 1990, 1994; Oakes, 1986; Rozeboom, 1960) consider that NHST and statistical inference have different aims. Regarding statistical inference, an attempt is made to know the probability of  $H_0$  being true considering the results and the data obtained in the sample ( $p(H_0|D)$ ). However, NHST provides information only about the probability of obtaining some equally or more discrepant data than what would be obtained if the  $H_0$  were true ( $p(D|H_0)$ ).

One of the most energetic defences against this criticism is the one presented by Hagen (1997). As a starting point for his argument, Hagen uses the article by Cohen (1994), which represents one of the strongest attacks on the significance test. The author considers that in the example proposed by Cohen to uphold his criticism, an attempt is made to associate the probability of  $H_0$  with relative frequencies of an empirical basis and susceptible to quantification, in such a way that  $H_0$  and  $H_1$  are understood as statements with respect to the sample. However, statistical hypotheses must always refer to the population (Hayes, 1963). In short, what Hagen proposes is that if we start from the conception of probability based on 'relative frequencies' (Fisher, 1937), the significance test does not provide the information expected by the researcher, but if the probability of  $H_0$  is put on a level with the 'degree of subjective belief' (Jeffrey, 1934), the aforementioned test provides the information it is actually intended to obtain.

Cortina and Dunlap (1997) point out that this criticism is not a deficiency of the significance test in itself, but rather an interpretation problem with regard to the information provided by the test in question.

#### *Logical validity of probabilistic nature premises*

This criticism, which is shared by the authors mentioned in the above section, among others, posits that the significance test is based on an incorrect use of syllogistic deductive reasoning and, specifically, of the rule known as 'Modus Tollens', due to the fact that the probabilistic statements are incompatible with the rules of deductive reasoning. This type of faulty reasoning is known as «the illusion of attaining improbability» (Falk & Greenbaum, 1995), and it is associated with several misconceptions, among which the following beliefs stand out: that the  $p$  value is the probability of  $H_0$  being true; that the probability of  $H_1$  being true is expressed by the complementary value of  $p$  ( $1-p$ ); that statistically significant results will be obtained if the experiment is replicated.

In response to this criticism, Cortina and Dunlap (1997) prove that the sequence of premises on which the significance test is based does not break the rules of syllogistic reasoning in those cases, common in psychology, in which the truthfulness of the antecedent and the consequent of the first premise are related in a positive way. From a different perspective, Hagen (1997) points out that a formally valid argument is not always appropriate and that, on the contrary, an argument can be reasonable and sustainable even when it does not have logical validity in a formal sense.

#### *Verification of psychological theories and advancement of scientific knowledge*

Cohen (1994) and Rozeboom (1960), among others, argue that the dichotomous decision to reject or accept  $H_0$  does not make it possible to test a psychological theory, since they consider it wrong to believe that the significance test allows us to know about the probability that the research hypothesis is true and, consequently, that the theory behind it has been confirmed. Other authors, like Carver (1978), Erwin (1998), Nickerson (2000) and Snow (1998), add that even when an  $H_0$  is objectively rejected, it is necessary to exclude another series of competing alternative hypotheses prior to verifying the validity of the research hypothesis. Thus, the increased truthfulness of this hypothesis can only be derived from a solid theoretical base, an appropriate research design and multiple replications of the study under different conditions.

Furthermore, many authors point out that NHST also fails to provide information about the practical importance and the magnitude of the observed effects (for example, Bakan, 1966; Bracey, 1991; Cohen, 1994; Meehl, 1997; Nickerson, 2000; Rosenthal, 1983; Shaver, 1985, 1993; Thompson, 1996).

With regard to this criticism, Chow (1996) argues that it is based on a clear confusion between statistical and inductive inference issues, adding that no statistical test enables one to prove a theory, since this verification is a process that goes beyond the refutation of a statistical hypothesis. However, objecting to the essence of this criticism, Mayo (1996) defends that the ability to make successful inductions is explained by the error statistical properties of our methods. Thus scientific knowledge grows because we have methods, such as NHST, that are adequate for learning from errors.

#### *Replicability of results*

Another criticism raised by several authors (Cohen, 1994; Falk & Greenbaum, 1995; Oakes, 1986; Shaver, 1993; Thompson, 1996) is the one referring to the false belief that the complementary value of  $p$  ( $1-p$ ) expresses the probability of the results being replicable («the replication deceit»). This would be true if the calculated  $p$  allowed us to know what the probability of  $H_0$  being true in the population was. However, as has been pointed out before, the calculated  $p$  does not provide this information.

Actually, this criticism is only valid when the  $H_0$  is true, since if the  $H_0$  is false and the effect size and the sample size of the replication coincide exactly with those of the original study, there is an increasing monotonic relationship between replicability and  $p$  values (Greenwald et al., 1996). On the other hand, Chow (1996) thinks that the nature of the sample distribution of the statistic, the

mathematical basis of the significance test, shows that there is nothing inherent to the significance test that leads to the erroneous interpretation underlying the replication deceit.

In a recent work, Killeen (2005) has proposed a new statistic, called  $p_{\text{rep}}$ , that estimates the probability of replicating an effect. Unlike  $p$ ,  $p_{\text{rep}}$  is not conditional on the truth value of the null, so it avoids reaching conclusions that may be logically consistent but empirically invalid.

#### *Informative usefulness*

Another important criticism levelled at the significance test is that it is not useful, due to the fact that the  $H_0$  does not provide substantive information and that the hypothesis mentioned is always false in the population (for example, Binder, 1963; Cohen, 1990, 1994; Meehl, 1997; Murphy, 1990; Oakes, 1986; Pollard, 1993; Thompson, 1992; Tukey, 1991; Weitzman, 1984). Consequently, the decision to reject it shows that the research design is powerful enough to detect an existing effect, regardless of its magnitude and usefulness. Given that, based on this fact, obtaining a statistically significant result depends more on the sample size than on the truthfulness or falsity of the research hypothesis or the appropriateness of the theory on which it is based (Hays, 1994; Oakes, 1986), several authors (Cohen, 1990, 1994; Kirk, 1996; Nickerson, 2000) suggest that it is ironic to be so concerned with controlling Type I Error and so little with controlling Type II Error (accepting levels as high as 0.5 or 0.8), whose consequences, in many applied contexts, are much more serious than those resulting from Type I Error. Agreeing with the premise that the  $H_0$  is always false, Cohen (1994) and Meehl (1997) conclude that the use of NHST is only valid in true experiments that include randomisation, or when the slightest deviation from pure chance can be important.

To refute this criticism, Hagen (1997) points out that the  $H_0$  does not propose equality among samples. Instead, supposing that there are certain differences, it postulates that samples have been obtained from the same population. And he adds that, contrary to what the supporters of this criticism think, when samples come from 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), 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. In the same line, Mayo (1996) also states that it is useful to test a null hypothesis asserting a zero difference in means—even knowing that the null is strictly false—because it teaches the extent to which an effect differs from zero.

Objecting directly to the essence of this criticism, Baril and Cannon (1995) and Frick (1995) point out that the  $H_0$  may be true, particularly in experiments in which the theory is of great importance and only one variable is handled, although they admit

that it is very difficult for the  $H_0$  to be true in experiments of a purely applied nature in which complex variables are manipulated.

#### *Dichotomous decision to reject/not reject $H_0$*

NHST has also been criticized because, when adopting a permanent significance level, the researcher turns an uncertainty continuum ranging from probability 0 to probability 1 into a dichotomous decision to reject/not reject the  $H_0$ . Furthermore, the criterion used for choosing the significance level that establishes the cut-off point for deciding whether the results are statistically significant or not is completely arbitrary (Glass, McGraw, & Smith, 1981; Johnson, 1999; Rosnow & Rosenthal, 1989; Rozeboom, 1960).

Contrary to the authors who criticize the arbitrariness in the election of the alpha value, Cox (1977) and Frick (1996) consider that this criterion has been appropriately established by the scientific community, which makes it possible to eliminate the influence of researchers' judgments and opinions on data interpretation, thus guaranteeing objectivity. Likewise, Chow (1996) thinks that it is an objective criterion at a mathematical level, whose meaning is not associated with the researcher's theoretical knowledge. Moreover, in the scientific community there is certain agreement about the arbitrariness in the election of the alpha value. However, it is worth pointing out that, based on an analysis of the history of statistical theory and probability, Cowles and Davis (1982) state that this election was not arbitrary, but rather derived from scientific conventions focused on the notion of chance and the improbability of the occurrence of a certain event.

#### *Main alternatives to NHST*

In order to overcome the problems related to NHST, several alternatives have been suggested, which, from the point of view of the most radical critics, should replace NHST and, in the view of the moderate ones, could be a complement to the significance test. The present article deals basically with the alternatives that, in this regard, are recommended by the TFSI of the APA (Wilkinson & TFSI, 1999) in the interest of improving the working methods of researchers in the framework of statistical analysis and data interpretation. Thus, for example, although the alternative based on Bayesian statistics could be an outstanding alternative to NHST, this approach is not dealt with 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. The following section explains the way the alternatives recommended by the TFSI can be calculated by means of SPSS statistical software, taking the data from a random factorial design as reference.

#### *Confidence intervals around point estimates*

The TFSI points out that «it is hard to imagine a situation in which a dichotomous accept-reject decision is better than providing an actual  $p$  value, or even better, a confidence interval» (p. 599).

Many authors consider that the calculation of confidence intervals around estimates is an excellent complement, or even substitute for, significance tests (for example, Brandstätter, 1999;

Hunter, 1997; Kirk, 1996, 2001; Loftus, 1991, 1995, 1996; Schmidt, 1996; Steiger & Fouladi, 1997; Thompson & Baugh, 2002; Valera, Sánchez, & Marín, 2000). These authors argue that, in addition to the information included in a statistical significance test, a confidence interval provides a ranking of values within which the actual population parameter is included, with a certain probability level. Thus, not only does it provide information about the nil-null hypothesis, but it also provides information about the  $H_0$  that do not take the zero value ('non-nil null hypothesis'). Moreover, the confidence interval reflects the accuracy in the population parameter estimate. In the case of confidence intervals for differences between parameters, these intervals not only make it possible to reject the hypothesis of no difference when the interval does not include zero, but they also indicate the direction and the magnitude of the difference in question. Furthermore, a point estimate and a confidence interval are governed by the same unit of measurement as the data, which aids in the interpretation of the results. Two other noteworthy advantages of confidence intervals, with regard to the significance tests, are the following: they enable the real error level to be maintained at 0.05 (or at the level established according to the given confidence interval); and they provide very useful information for carrying out meta-analytic studies.

Summarizing its advantages, confidence intervals avoid many of the problems inherent to the significance test. They do not require a hypothesis to be formulated a priori, nor do they test trivial hypotheses. Moreover, they provide a greater amount of information, and they are easier to interpret than the significance test. However, as Estes (1997) states, confidence intervals are also subject to misconstruals and misuses. A property of them that may interfere with efficient communication of information is the inherent association of a confidence interval with a particular choice of significance level. This implies that a single significance level should be adopted for all tests performed within any one research report to make the same interpretation of the confidence intervals exhibited in different figures because, otherwise, conflicting connotations adds to the processing demand on the reader. In the interest of consistency and unambiguity of communication, Estes suggests to exhibit only actual confidence intervals in figures and to do so in connection with the issue of how closely population means are estimated by obtained sample means.

### *Effect sizes*

The TFSI recommends presenting, in all cases, effect sizes for the main results. Moreover, it places special emphasis on the need for interpreting effect sizes within a practical and theoretical context, and it highlights the importance of these indices in carrying out power analyses and meta-analyses in the future (p. 599).

Many authors think effect sizes should be calculated and interpreted whenever research is carried out (e.g., Brandstätter, 1999; Carver, 1978; 1993; Cohen, 1988; Cook & Campbell, 1979; Fisher, 1925; Folger, 1989; Glass, 1976; Harris, 1991; Kirk, 1996, 2001; Rosenthal, 1984; Rosnow & Rosenthal, 1989; Schmidt, 1996; Snyder & Lawson, 1993).

However, numerous studies have shown that, in the field of Psychology, the effect size is not sufficiently taken into account (Chase & Chase, 1976; Clark-Carter, 1997; Frías, García, &

Pascual, 1994; Rossi, 1990; Sánchez, Valera, Velandrino, & Marín, 1992; Seldmeier & Gigerenzer, 1989; Valera, Sánchez, & Marín, 2000), and that there are deficiencies in its interpretation (Monterde, Pascual, & Frías, 2006).

In addition to providing information about the magnitude of the observed effect, the effect size makes it possible to directly compare the results obtained in different studies, since these indices are transformations onto a common scale. They are also an essential element for performing power analyses within the framework of statistical significance tests and carrying out meta-analytic studies. Furthermore, when the measurement scales of the variables are unfamiliar, the standardized measurements of the effect sizes and their confidence intervals can provide information about the practical significance of the results obtained in a certain study.

### *Confidence intervals for effect sizes*

The TFSI recommends providing confidence intervals for all effect sizes associated with the main results. It also highlights the importance of comparing confidence intervals across different studies, instead of merely verifying whether such intervals include the zero value. Furthermore, it warns about the common error of assuming that a parameter is included within a confidence interval (p.599).

An increasing number of authors follow the TFSI recommendation that refers to calculating confidence intervals for effect size measures (for example, Cumming & Finch, 2001; Fidler & Thompson, 2001; Robinson & Wainer, 2001; Schmidt, 1996; Smithson, 2001; Thompson, 2002). They do so because these intervals provide easily understandable information that helps to interpret the results appropriately, they are highly useful in gathering empirical evidence across different studies, and their breadth provides information about the accuracy of the estimate. Moreover, there is an association between confidence intervals and NHST, which can lead to a better understanding of the logic underlying both strategies.

### *Power analysis*

Since, according to the TFSI, power analysis makes more sense when carried out before gathering and examining data, this organism recommends calculating a power analysis ranking in order to observe how power estimates change depending on different effect sizes and alpha levels. It also suggests that, in the description of the results, the calculated power should be replaced by confidence intervals (p. 596).

This analysis is considered to be especially relevant when, after the impossibility of rejecting the  $H_0$ , the intention is to conclude that there is no effect or that, if it does exist, its magnitude is insignificant (Meehl, 1991; Robinson & Levin, 1997; Schafer, 1993).

Cohen's findings, with respect to the limited power of the majority of studies carried out in psychology (Cohen, 1962), are still quite applicable nowadays (Kazdin & Bass, 1989; Rosnow & Rosenthal, 1989; Seldmeier & Gigerenzer, 1989; Valera, Sánchez, & Marín, 2000). This fact, according to detractors of NHST, is a serious problem that hinders the advancement of knowledge, and it is closely associated with the erroneous interpretations made about the information provided by the significance test.

### Replication

The TFSI warns that, in their desire to reject the  $H_0$ , researchers can make the mistake of publishing false theories derived from the use of an inappropriate methodology, even when the statistical analysis is correct. It also suggests carrying out replications of the original study in order to avoid this problem (p. 600).

As Allen and Preiss (1993) state, scientific knowledge is reached by means of replication. The results from an unreplicated study, regardless of the statistical significance achieved, are bound to be speculative (Hubbard & Armstrong, 1994) and do not make sense by themselves (Lindsay & Ehrenberg, 1993).

#### Estimate of alternative procedures to the significance test in between subjects designs with SPSS

As we have already pointed out in the introduction, after succinctly describing the current situation of the controversy surrounding the significance test in psychology, we will try to lay the foundations for broadening the debate beyond the methodological field. With this aim in mind, we will show the way data are analysed within the framework of the significance test. We will also calculate, using SPSS, the main indices proposed to complement (or replace) the test under examination, starting with data from a design widely used in experimental research, the two-way between factorial design (ANOVA model). It is important to point out that the procedures described here can be generalized to any between-subjects factorial design. The experimental designs that incorporate repeated measures factors require modifications in some of these procedures, mainly with regard to the error terms for the effect size indices (see Olejnik & Algina, 2000).

As an example, let us suppose that a researcher decides to carry out a study in which he manipulates the role adopted by the subject (social role/individual role) and the type of information provided (statistical/anecdotal), in order to examine the influence of these factors on the precautionary behaviour adopted when driving. With the aim of carrying out the research, a sample consisting of 40 subjects is gathered, and 10 people are randomly assigned to each of the four experimental conditions. After measuring the amount of precautionary behaviour shown by the subject when driving, the results are shown in table 1.

| Table 1<br>Amount of precautionary behaviour shown by the subject |                       |                         |                       |
|-------------------------------------------------------------------|-----------------------|-------------------------|-----------------------|
| Social role                                                       |                       | Individual role         |                       |
| Statistical information                                           | Anecdotal information | Statistical information | Anecdotal information |
| 7                                                                 | 3                     | 2                       | 5                     |
| 6                                                                 | 1                     | 1                       | 7                     |
| 8                                                                 | 2                     | 2                       | 6                     |
| 10                                                                | 3                     | 3                       | 9                     |
| 5                                                                 | 3                     | 1                       | 5                     |
| 6                                                                 | 4                     | 4                       | 8                     |
| 9                                                                 | 1                     | 2                       | 5                     |
| 8                                                                 | 5                     | 5                       | 4                     |
| 7                                                                 | 3                     | 2                       | 7                     |
| 6                                                                 | 1                     | 3                       | 5                     |

### Null hypothesis significance testing

The analytic model typically used in order to carry out NHST in two-way factorial designs is the two-way analysis of variance, whose syntax, based on the modulus *General Linear Model (Univariate)* of SPSS statistical software, is the following:

#### UNIANOVA

Behaviour BY Information Role

/METHOD = SSTYPE(3)

/INTERCEPT = INCLUDE

/PRINT = DESCRIPTIVE ETASQ OPOWER PARAMETER (Syntax 1)

/CRITERIA = ALPHA(.05)

/DESIGN = Information Role Information\*Role.

In addition to calculating the indices associated with NHST, which are obtained by default, descriptive statistics, estimates of effect size, observed power and parameter estimates are obtained by means of this syntax, issues which we will deal with in different epigraphs of the present study.

As can be observed in figure 1, the results obtained indicate the existence of a statistically significant interaction ( $F(1,36) = 80.26$ ;  $p = 0.0001$ ) between the type of role and type of information variables, with 67.3% of the total variance being explained by the proposed model ( $R^2_{\text{adjusted}} = 0.673$ ).

One of the recommended strategies for examining the interaction effect consists of studying the simple effects. Although SPSS software does not include a specific command for calculating these effects, this analysis can be carried out by means of a straightforward modification of the syntax (see syntax 2). Among the different options provided by the SPSS for performing the confidence interval adjustment, we have chosen the Bonferroni adjustment for controlling the type I error rate when estimating simple effects.

#### UNIANOVA

Behaviour BY Role Information

/METHOD = SSTYPE(3)

(Syntax 2)

/INTERCEPT = INCLUDE

/EMMEANS=TABLES(Role\*Information)COMPARE(Information)ADJ(BONFERRONI)

/EMMEANS = TABLES(Information) COMPARE

ADJ(BONFERRONI)

/CRITERIA = ALPHA(.05)

/DESIGN = Role Information Role\*Information.

The results obtained are illustrated in figure 2.

### Confidence intervals around point estimates

The option concerning parameter estimates makes it possible to obtain confidence intervals around point estimates (see figure 3). Figure 3 provides partial regression coefficients,  $p$  significance levels of empirical results and confidence intervals around point estimates.

### Effect sizes

The analysis of variance provides the value of the adjusted  $R^2$  as the variance proportion explained by the proposed model.

Nevertheless, the magnitude of the effect in each contrast, in terms of proportion of explained variance, can be calculated by means of eta squared (Pearson, 1905), epsilon squared (Kelley, 1935) and omega squared (Hays, 1963) coefficients, among others. Eta squared is a biased estimator that tends to overvalue the intensity of the relationship. Epsilon squared corrects the numerator of eta squared by subtracting the error mean square from the contrast sum of squares. Finally, omega squared adjusts epsilon squared by adding the error mean square to the total sum of squares in the denominator of epsilon squared.

In the example we are concerned with here, the value of partial eta squared for the interaction effect is 0.69 (see figure 1). Partial eta squared is an estimator supplied by SPSS program that is defined as

the proportion of total variation attributable to the factor, partialling out (excluding) other factors from the total nonerror variation. Since eta squared is a biased estimator of the effect size, it is better to calculate the partial or/and total value of omega squared. In the total omega squared coefficient, the variance due to the effect we are interested in is expressed as a proportion of the sum of the error variance and the variances of the effects of all factors involved in the design. Its calculation, for a fixed effects factorial design, is carried out by means of the following expression:

$$\hat{\omega}_{total}^2 = \frac{df_{effect}(MS_{effect} - MS_{error})}{SS_{total} + MS_{error}} \quad (1)$$

| Tests of between-subjects effects |                         |    |             |         |      |                     |                    |                             |
|-----------------------------------|-------------------------|----|-------------|---------|------|---------------------|--------------------|-----------------------------|
| Source                            | Type III Sum of squares | df | Mean square | F       | Sig. | Partial eta squared | Noncent. Parameter | Observed Power <sup>a</sup> |
| Corrected model                   | 174,200 <sup>b</sup>    | 3  | 58,067      | 27,724  | ,000 | ,698                | 83,172             | 1,000                       |
| Intercept                         | 846,400                 | 1  | 846,400     | 404,117 | ,000 | ,918                | 404,117            | 1,000                       |
| Information                       | 2,500                   | 1  | 2,500       | 1,194   | ,282 | ,032                | 1,194              | ,186                        |
| Role                              | 3,600                   | 1  | 3,600       | 1,719   | ,198 | ,046                | 1,719              | ,248                        |
| Information * role                | 168,100                 | 1  | 168,100     | 80,260  | ,000 | ,690                | 80,260             | 1,000                       |
| Error                             | 75,400                  | 36 | 2,094       |         |      |                     |                    |                             |
| Total                             | 1096,000                | 40 |             |         |      |                     |                    |                             |
| Corrected total                   | 249,600                 | 39 |             |         |      |                     |                    |                             |

<sup>a</sup> Computed using alpha= ,05  
<sup>b</sup> R Squared= ,698 (Adjusted R Squared= ,673)

Figure 1. Two-way analysis of variance

| Pairwise comparisons                                                                                      |                         |                         |                       |             |                   |                                                     |             |
|-----------------------------------------------------------------------------------------------------------|-------------------------|-------------------------|-----------------------|-------------|-------------------|-----------------------------------------------------|-------------|
| Dependent variable: Behavior                                                                              |                         |                         |                       |             |                   |                                                     |             |
|                                                                                                           |                         |                         |                       |             |                   | 95% Confidence interval for difference <sup>a</sup> |             |
| Role                                                                                                      | (I) Information         | (J) Information         | Mean difference (I-J) | Std. error  | Sig. <sup>a</sup> | Lower bound                                         | Upper bound |
| Social role                                                                                               | Statistical information | Anecdotal information   | 4,600*                | ,647        | ,000              | 3,287                                               | 5,913       |
|                                                                                                           | Anecdotal information   | Statistical information | -4,600*               | ,647        | ,000              | -5,913                                              | -3,287      |
| Individual role                                                                                           | Statistical information | Anecdotal information   | -3,600*               | ,647        | ,000              | -4,913                                              | -2,287      |
|                                                                                                           | Anecdotal information   | Statistical information | 3,600*                | ,647        | ,000              | 2,287                                               | 4,913       |
| Based on estimated marginal means                                                                         |                         |                         |                       |             |                   |                                                     |             |
| * The mean difference is significant at the ,05 level.                                                    |                         |                         |                       |             |                   |                                                     |             |
| <sup>a</sup> Adjustment for multiple comparisons: Bonferroni.                                             |                         |                         |                       |             |                   |                                                     |             |
| Univariate tests                                                                                          |                         |                         |                       |             |                   |                                                     |             |
| Dependent variable: Behavior                                                                              |                         |                         |                       |             |                   |                                                     |             |
| Role                                                                                                      |                         | Sum of squares          | df                    | Mean square | F                 | Sig.                                                |             |
| Social role                                                                                               | Contrast                | 105,800                 | 1                     | 105,800     | 50,515            | ,000                                                |             |
|                                                                                                           | Error                   | 75,400                  | 36                    | 2,094       |                   |                                                     |             |
| Individual role                                                                                           | Contrast                | 64,800                  | 1                     | 64,800      | 30,939            | ,000                                                |             |
|                                                                                                           | Error                   | 75,400                  | 36                    | 2,094       |                   |                                                     |             |
| Each F examines the simple effects of information within each level of Role.                              |                         |                         |                       |             |                   |                                                     |             |
| These tests are based on the linearly independent pairwise comparisons among the estimated marginal means |                         |                         |                       |             |                   |                                                     |             |

Figure 2. Simple effects estimates

| Parameter estimates                  |                |            |        |      |             |             |                     |                    |                             |
|--------------------------------------|----------------|------------|--------|------|-------------|-------------|---------------------|--------------------|-----------------------------|
| Dependent variable: Behavior         |                |            |        |      |             |             |                     |                    |                             |
| 95% Confidence interval              |                |            |        |      |             |             |                     |                    |                             |
| Parameter                            | B              | Std. error | t      | Sig. | Lower bound | Upper bound | Partial eta squared | Noncent. parameter | Observed Power <sup>a</sup> |
| Intercept                            | 6,100          | ,458       | 13,33  | ,000 | 5,172       | 7,028       | ,832                | 13,329             | 1,000                       |
| [Information= 1,00]                  | -3,600         | ,647       | -5,562 | ,000 | -4,913      | -2,287      | ,462                | 5,562              | 1,000                       |
| [Information= 2,00]                  | 0 <sup>b</sup> | –          | –      | –    | –           | –           | –                   | –                  | –                           |
| [Role= 1,00]                         | -3,500         | ,647       | -5,408 | ,000 | -4,813      | -2,187      | ,448                | 5,408              | 1,000                       |
| [Role= 2,00]                         | 0 <sup>b</sup> | –          | –      | –    | –           | –           | –                   | –                  | –                           |
| [Information= 1,00]*<br>[Role= 1,00] | 8,200          | ,915       | 8,959  | ,000 | 6,344       | 10,056      | ,690                | 8,959              | 1,000                       |
| [Information= 1,00]*<br>[Role= 2,00] | 0 <sup>b</sup> | –          | –      | –    | –           | –           | –                   | –                  | –                           |
| [Information= 2,00]*<br>[Role= 1,00] | 0 <sup>b</sup> | –          | –      | –    | –           | –           | –                   | –                  | –                           |
| [Information= 2,00]*<br>[Role= 2,00] | 0 <sup>b</sup> | –          | –      | –    | –           | –           | –                   | –                  | –                           |

<sup>a</sup> Computed using alpha= .05  
<sup>b</sup> This parameter is set to zero because it is redundant

Figure 3. Parameter estimates

On the other hand, in a partial variance proportion, the variance due to the effect we are interested in is expressed as a proportion of the sum of the error variance and the variance of this effect, excluding the variance from the rest of the factors. Its calculation, for a fixed effects factorial design, is carried out by means of the following expression:

$$\hat{\omega}_{\text{partial}}^2 = \frac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{df_{\text{effect}}MS_{\text{effect}} + (N - df_{\text{effect}})MS_{\text{error}}} \quad (2)$$

With regard to the choice between the two indices, it must be taken into account that when using the total variance as standardizer, this includes all the design sources of variation, which makes it difficult to compare this coefficient to those obtained through other designs that incorporate different factors. Furthermore, the partial variance proportion reveals the problem that the effects of different factors included in the same design cannot be compared, due to the fact that they do not share the same denominator (standardizer) (Olejnik & Algina, 2000).

Starting from the data obtained in figure 1, the values of total and partial omega squared for the interaction effect are 0.659 and 0.66, respectively. Following the criterion proposed by Cohen (1988) for the interpretation of effect sizes, the magnitude of the effect associated with the interaction is high.

Nevertheless, it is important to point out that, in the case concerning us, the information provided by the variance proportion as effect size associated with the interaction is, from any point of view, incomplete. Thus, it is essential to estimate the effect size for each of the previously calculated simple effects.

In this regard, it must be taken into account that, in order to estimate the effect size associated with a certain contrast, the standardized difference among means is more frequently used than the variance proportion (Olejnik & Algina, 2000), with *Cohen's d*,

*Glass' g'* and *Hedges' g* being the most widely used indices in the realm of psychological research (see formulas 3, 4, 5 and 6). The differences between the three indices lie in the way standardization is carried out, and authors like Huberty (2002) and Olejnik and Algina (2000) consider that *Hedges' g* is the most suitable when variances of *J* groups involved in the contrast estimate the same population variance. If the homoscedasticity assumption is not met, these authors suggest that the standard deviation of one of the groups should be used.

$$d_{\text{Cohen}} = \frac{\bar{x}_1 - \bar{x}_2}{S_{\text{pooled\_comp}}} \quad (3)$$

where  $S_{\text{pooled\_comp}}$  = Pooled standard deviation of the groups that are being compared.

$$g'_{\text{Glass}} = \frac{\bar{x}_E - \bar{x}_C}{S_C} \quad (4)$$

where  $S_C$  = Standard deviation of one of the groups, usually the control group.

$$g_{\text{Hedges}} = \frac{\bar{x}_1 - \bar{x}_2}{S_{\text{pooled\_total}}} \quad (5)$$

where  $S_{\text{pooled\_total}}$  = Pooled standard deviation of all the groups involved in the design.

$$S_{\text{pooled}}^2 = \frac{(n_1 - 1)S_1^2 + \dots + (n_j - 1)S_j^2}{(n_1 - 1) + \dots + (n_j - 1)} \quad (6)$$

Next, we will go on to calculate the value of Cohen's  $d$ , given the descriptive statistics obtained in the analysis.

The first estimated simple effect (see figure 2) corresponds to the difference between the precautionary behaviour of subjects who adopt the social role when receiving statistical information or when receiving anecdotal information, with the value of Cohen's  $d$  being 3.17 for this contrast. The second simple effect corresponds to the difference between the precautionary behaviour adopted by individual decision-makers depending on whether the information received is statistical or anecdotal, with a 2.1 value of Cohen's  $d$  being obtained for this contrast.

In order to interpret this index, Cohen (1988) suggests the 0.20, 0.50 and 0.80 values as low, medium and high values of effect sizes, respectively. Based on this reference, the values of the effect sizes obtained in our example are very high.

#### Confidence intervals for effect sizes

In contrast to the type of distributions (central distributions) used in order to calculate confidence intervals for means, correlation coefficients and other statistics, like Student's  $t$  or Fisher's  $F$ , the estimate of confidence intervals for effect sizes requires the use of non-central distributions. Currently used statistical software (except *Statistica* software, which includes an additional module for the calculation of confidence intervals based on non-central distributions) does not include the calculation of the confidence interval for effect size indices. To this end, specific routines created for statistical software or some type of specialized software must be used (for example, Steiger and Fouladi's  $R^2$  for the calculation of confidence intervals of  $R^2$ : [www.interchg.ubc.ca/steiger/homepage.htm](http://www.interchg.ubc.ca/steiger/homepage.htm), or G. Cumming's ESCI: <http://www.latrobe.edu.au/psy/esci/>).

With the aim of calculating the confidence interval for eta squared associated with the interaction effect, we must start from the results obtained in the analysis of variance (see figure 1), and we must insert the following in the data editor of the SPSS: the value obtained in the  $F$  test ( $f_{val}$ ), the degrees of freedom of the source of variation ( $df_1$ ) and of the error ( $df_2$ ), the confidence level ( $conf$ ) and the corrected total sum of squares ( $sostot$ ) (see figure 4).

Afterward, we must perform the syntax for SPSS proposed by Smithson for non-central  $F$  distributions (NoncF3). This syntax is

not included in this article, due to space restrictions, but it can be downloaded from the following website: <http://www.anu.edu.au/psychology/people/smithson/details/CIstuff/CI.html>.

After performing the syntax, we obtain the value of eta squared ( $R^2$ ), as well as lower ( $LR^2$ ) and upper ( $UR^2$ ) limits of the confidence interval for eta squared (see figure 4).

Furthermore, this figure provides the limits of the confidence interval ( $LC^2$  and  $UC^2$ ), the non-centrality parameter and the power of the test ( $POWER$ ).

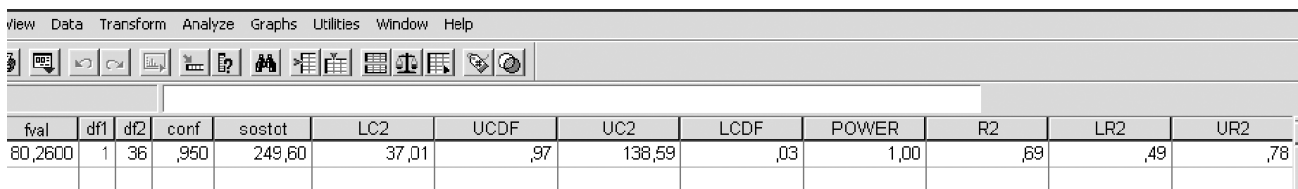
In order to calculate the non-central confidence interval for omega squared, limits of the CI for eta squared can be transformed by means of syntax 3, as proposed by Fidler and Thompson (2001, p. 593).

```
compute sosb_L=sostot*lr2.
execute.
compute msw_L=(sostot - sosb_L)/df2.
execute.
compute omega_L=(sosb_L-(df1*msw_L))/(sostot+msw_L).
execute.
compute sosb_U=sostot*ur2.
execute.
compute msw_U=(sostot - sosb_U)/df2.
compute omega_U=(sosb_U-(df1*msw_U))/(sostot+msw_U).
execute.
```

(Syntax 3)

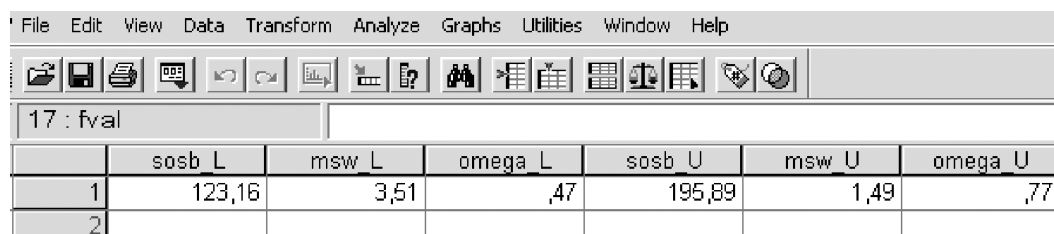
After carrying out the syntax, we obtain the values of lower ( $\omega_L$ ) and upper ( $\omega_U$ ) limits of the confidence interval for omega squared (see figure 5).

As in the case of the proportion of explained variance, the calculation of the 95% confidence interval for standardized differences among means requires the use of a non-central  $t$  distribution. In the example in question, confidence intervals for effect sizes obtained in the analysis of simple effects have been calculated. To this end, the value obtained in Student's  $t$  test ( $t_{val}$ ), the degrees of freedom ( $df$ ), the confidence level ( $conf$ ) and the sizes of the groups ( $n_1$  and  $n_2$ , respectively) must be inserted for each contrast in a data file of SPSS (see figure 6). Afterward, the syntax (NoncT2 and T2D) for SPSS proposed by Smithson for the calculation of confidence



| fval    | df1 | df2 | conf | sostot | LC2   | UCDF | UC2    | LCDF | POWER | R2  | LR2 | UR2 |
|---------|-----|-----|------|--------|-------|------|--------|------|-------|-----|-----|-----|
| 80,2600 | 1   | 36  | ,950 | 249,60 | 37,01 | ,97  | 138,59 | ,03  | 1,00  | ,69 | ,49 | ,78 |

Figure 4. Data that must be inserted and results obtained in the calculation of the Confidence Interval for eta squared



|   | sosb_L | msw_L | omega_L | sosb_U | msw_U | omega_U |
|---|--------|-------|---------|--------|-------|---------|
| 1 | 123,16 | 3,51  | ,47     | 195,89 | 1,49  | ,77     |
| 2 |        |       |         |        |       |         |

Figure 5. Confidence Interval for omega squared



intervals based on non-central  $t$  distributions must be carried out (these files can be downloaded from the following website: <http://www.anu.edu.au/psychology/people/smithson/details/CIstuff/CI.html>).

In addition to the values associated with the non-centrality parameter (LC2, UCDF, UC2 and LCDF) and the power of the test (POWER), the results obtained (see figure 6) show the limits of the confidence interval (LOWD and HIGHD, respectively) for the standardized differences among means.

#### Power analysis

As can be observed in the different results figures shown in the present article, the power of the test is included in the analysis of variance (figures 1 and 3) and the figures on the calculation of confidence intervals for the effect size (figures 4 and 6). In the example we are using, this value is 1.00 for the estimate of the interaction effect, as well as for the calculation of confidence intervals for omega squared and Cohen's  $d$ .

However, it is worth pointing out that power analysis makes more sense when carried out before gathering and examining data than when performed a posteriori.

#### Conclusions

Summarizing the controversy presented in the first section of this article, we believe that the alternative indices examined in the present paper should compliment rather than replace the information provided by the significance test. The goal is to increase the rigor in the assessment of results. There is a need to

make researchers understand that the significance test only provides information about whether or not there is a rational basis for excluding sample error as a plausible explanation for the data. And this is the step prior to the process of searching for specific non-random factors able to explain such data. It is important for researchers to become aware of the fact that the use and scope of the significance test should be restricted to the merely statistical field. Therefore, we consider that to increase the quality in data analysis and interpretation, there is a need to use the significance test along with other procedures that make it possible to obtain different types of information from the one provided by this index. We defend the eradication of dogmatic attitudes and the systematic use of statistical recipes, and we support the reflective use of the available procedures to fully evaluate the results of psychological research.

As several authors point out (Fidler, 2002; Hubbard & Ryan, 2000; Kirk, 1996; Robinson & Wainer, 2001; Schmidt, 1996), the rejection of the exclusive use of the significance test and the acquisition of new habits in data analysis and interpretation can only be achieved by tackling the problem from multiple contexts. In order to reach this goal, it is necessary to obtain the involvement of groups, such as those that include text book authors, lecturers teaching methodological subjects in undergraduate and graduate courses, authors of doctoral theses, creators of statistical software, editors and reviewers of scientific journals, and bodies producing manuals on publication guidelines for scientific works. We consider that, until such involvement takes place, providing researchers with the tools that allow them to approach the existing controversy surrounding the significance test from a practical perspective is a very useful strategy for moving forward.

|   | tval   | df | conf | n1    | n2    | LC2  | UCDF | UC2  | LCDF | POWER | LOWD | HIGHD |
|---|--------|----|------|-------|-------|------|------|------|------|-------|------|-------|
| 1 | 7,1070 | 36 | ,950 | 20,00 | 20,00 | 4,53 | ,97  | 9,63 | ,03  | 1,00  | 1,43 | 3,05  |
| 2 | 5,5600 | 36 | ,950 | 20,00 | 20,00 | 3,19 | ,97  | 7,87 | ,02  | 1,00  | 1,01 | 2,49  |
| 3 |        |    |      |       |       |      |      |      |      |       |      |       |

Figure 6. Data that must be inserted and results obtained in the calculation of the Confidence Interval for Cohen's  $d$

#### 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.
- Bakan, D. (1966). The tests of significance in psychological research. *Psychological Bulletin*, 66, 423-437.
- Balluerka, N., Gómez, J., & Hidalgo, M.D. (2005). The controversy over null hypothesis significance testing revisited. *METHODOLOGY. European Journal of Research Methods for the Behavioral and Social Sciences*, 1(2), 55-70.
- Baril, G.L., & Cannon, J.T. (1995). What is the probability that null hypothesis testing is meaningless? *American Psychologist*, 50, 1098-1099.
- Binder, A. (1963). Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, 70, 107-115.
- Brace, G.W. (1991). Sense, non-sense and statistics. *PhiDelta Kappan*, 73, 335.
- Branstatter, E. (1999). Confidence intervals as an alternative to significance testing. *Methods of Psychological Research Online*, 4(2), 33-46.
- Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review*, 48(3), 378-399.
- Carver, R.P. (1993). The case against statistical significance testing revisited. *Journal of Experimental Education*, 61, 287-292.
- Chase, L.J., & Chase, R.B. (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology*, 61, 234-237.
- 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.
- Clark-Carter, D. (1997). The account taken of statistical power in research published in the British Journal of Psychology. *British Journal of Psychology*, 88, 71-83.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145-153.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (rev. ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45(12), 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.
- Cortina, J.M., & Dunlap, W.P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, 2(2), 161-172.
- Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, 37(5), 553-558.
- Cox, D.R. (1977). The role of significance tests. *Scandinavian Journal of Statistics*, 4, 49-70.
- Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and non-central distributions. *Educational and Psychological Measurement*, 61, 532-574.
- Dixon, P. (1998). Why scientists value  $p$  values. *Psychonomic Bulletin and Review*, 5, 390-396.
- Erwin, E. (1998). The logic of null hypothesis testing. *Behavioral and Brain Sciences*, 21, 197-198.
- Estes, W.K. (1997). On the communication of information by displays of standard errors and confidence intervals. *Psychonomic Bulletin and Review*, 4(3), 330-341.
- 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(5), 749-770.
- Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed-and random-effects effect sizes. *Educational and Psychological Measurement*, 61(4), 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(2), 181-210.
- Fisher, R.A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.
- Fisher, R.A. (1937). *The design of experiments*. London: Oliver & Boyd.
- Folger, R. (1989). Significance tests and the duplicity of binary decisions. *Psychological Bulletin*, 106, 155-160.
- Frías, D., García, J.F., & Pascual, J. (1994). Estudio de la potencia de los trabajos publicados en Psicológica. Estimación del número de sujetos fijando  $\alpha$  y  $\beta$ . *III Simposium de Metodología de las Ciencias Sociales y del Comportamiento* (pp. 1057-1063). Santiago de Compostela: Servicio de Publicaciones de la Universidad de Santiago de Compostela.
- 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(4), 379-390.
- Glass, G.V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher*, 5, 3-8.
- Glass, G.V., McGraw, B., & Smith, M.L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- Grant, D.A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological review*, 69, 54-61.
- Greenwald, A.G., González, 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.
- Hagen, R.L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, 52(1), 15-24.
- Harris, R.J. (1991). Significance tests are not enough: The role of effect-size estimation in theory corroboration. *Theory and Psychology*, 1, 375-382.
- Hays, W.L. (1963). *Statistics for psychologists*. New York, NY: Holt, Rinehart & Winston.
- Hays, W.L. (1994). *Statistics* (4th ed.). New York: Holt, Rinehart and Winston.
- 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(4), 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(5), 661-681.
- Huberty, C.J. (1987). On statistical testing. *Educational Researcher*, 16(8), 4-9.
- Huberty, C.J. (2002). A history of effect size indices. *Educational and Psychological Measurement*, 62, 227-240.
- Hunter, J.E. (1997). Need: A ban on the significance test. *Psychological Science*, 8, 3-7.
- Jeffrey, 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.
- Kelley, T.L. (1935). An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences*, 21, 554-559.
- Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, 16(5), 345-353.
- Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56(5), 746-759.
- Kirk, R.E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, 61(2), 213-218.
- 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. (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.
- Mayo, D. (1996). *Error and the growth of experimental knowledge*. Chicago: The University of Chicago Press.
- 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. Wile (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.
- Monterde, H., Pascual, J., & Frías, M.D. (2006). Errores de interpretación de los métodos estadísticos: importancia y recomendaciones. *Psicothema*, 18(4), 848-856.
- Murphy, K.R. (1990). If the null hypothesis is impossible, why test it? *American Psychologist*, 45(3), 403-404.

- Nickerson, R.S. (2000). Null hypothesis significance testing: A review of and old and continuing controversy. *Psychological Methods*, 5(2), 241-301.
- Oakes, M. (1986). *Statistical inference: A commentary for social and behavioral sciences*. New York: Wiley.
- Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations and limitations. *Contemporary Educational Psychology*, 25, 241-286.
- Pearson, K. (1905). *Mathematical contributions to the theory of evolution: XIV. On the general theory of skew correlations and nonlinear regression* (Draper's Company Research Memoirs, Biometric Series II). London: Dulau.
- Pollard, P. (1993). How significant is «significance»? In G. Keren & C. Lewis (Eds.): *A handbook for data analysis in the behavioural sciences: Volume 1. Methodological issues*. Hillsdale, NJ: Erlbaum.
- 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.
- 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*, 5, 646-656.
- Rozeboom, W.W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, 57, 416-428.
- Sánchez, J., Valera, A., Velandrino, A.P., & Marín, F. (1992). Un estudio de la potencia estadística en Anales de Psicología. *Anales de Psicología*, 8, 19-32.
- Schafer, W.D. (1993). Interpreting statistical significance and non-significance. *Journal of Experimental Education*, 61, 383-387.
- Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1(2), 115-129.
- 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 behavioural sciences: Volume 1. Methodological Issues* (pp. 199-228). Hillsdale, NJ: Erlbaum.
- 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 it is not. *Journal of Experimental Education*, 61(4), 293-316.
- 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(4), 305-632.
- 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.
- Steiger, J.H., & Fouladi, R.T. (1997). Noncentral 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.
- 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. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, 25(2), 26-30.
- Thompson, B. (2002). «Statistical», «practical» and «clinical»: How many kinds of significance do counsellors need to consider? *Journal of Counseling and Development*, 80, 64-71.
- Thompson, B., & Baugh, F. (2002). Using effect sizes, confidence intervals and especially confidence intervals for effect sizes: New APA and journal mandates for improved practices. *Metodología de las Ciencias del Comportamiento, special number*, 539-543.
- Tukey, J.W. (1991). The philosophy of multiple comparisons. *Statistical Science*, 6, 100-116.
- Valera, A., Sánchez, J., & Marín, F. (2000). Contraste de hipótesis e investigación psicológica española: análisis y propuestas. *Psicothema*, 12(2), 549-552.
- Weitzman, R.A. (1984). Seven treacherous pitfalls of statistics, illustrated. *Psychological Reports*, 54(2), 355-363.
- Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.