Original Article

Inconsistencies in Reported *p*-Values in Spanish Journals of Psychology

The Case of Correlation Coefficients

José Manuel Caperos,[1] Ricardo Olmos,[2] and Antonio Pardo[2]

[1]Centro de Ciencias de la Salud San Rafael-Nebrija, Madrid, Spain
[2]Universidad Autónoma de Madrid, Spain

Abstract: Correlation analysis is one of the most widely used methods to test hypotheses in social and health sciences; however, its use is not completely error free. We have explored the frequency of inconsistencies between reported *p*-values and the associated test statistics in 186 papers published in four Spanish journals of psychology (1,950 correlation tests); we have also collected information about the use of one- versus two-tailed tests in the presence of directional hypotheses, and about the use of some kind of adjustment to control Type I errors due to simultaneous inference. Reported correlation tests (83.8%) are incomplete and 92.5% include an inexact *p*-value. Gross inconsistencies, which are liable to alter the statistical conclusions, appear in 4% of the reviewed tests, and 26.9% of the inconsistencies found were large enough to bias the results of a meta-analysis. The election of one-tailed tests and the use of adjustments to control the Type I error rate are negligible. We therefore urge authors, reviewers, and editorial boards to pay particular attention to this in order to prevent inconsistencies in statistical reports.

Keywords: correlation analysis, inconsistencies in *p*-values, meta-analysis, Type I error rate, one-tailed tests

The *null hypothesis significance testing* (NHST) has for decades and continues to be the most widespread method used to analyze data in social and health sciences. The method is applied in more than 95% of empirical papers published in psychology journals (see Cumming et al., 2007).

Regardless of the controversy concerning the utility and validity of NHST (a controversy that has accompanied this analytical strategy since its inception; see Balluerka, Gómez, & Hidalgo, 2005; Chow, 1996; Nikerson, 2000), it is common to find errors in research reports which are related to the way data are analyzed and interpreted (Bakker & Wicherts, 2014; Curran-Everett, 2000; Jeličić, Phelps, & Lerner, 2009; Pardo, Garrido, Ruiz, & San Martín, 2007; Rosnow & Rosenthal, 1989; Veldkamp, Nuijten, Dominguez-Alvarez, van Assen, & Wicherts, 2014). Furthermore, some reviews carried out in recent years in different fields of knowledge have revealed the presence of inconsistencies in the reported *p*-values (Bakker & Wicherts, 2011; Berle & Starcevic, 2007; Caperos & Pardo, 2013; García-Berthou & Alcaraz, 2004).

The American Psychological Association (APA) recommends including all data needed to assess the applied statistics in reports and it specifically recommends describing the value of the test statistic, the degrees of freedom, and the associated exact *p*-value (APA, 2010, p. 34). Several authors have questioned psychological research, emphasizing the importance of evaluating the quality of the studies and research reports in this field (Bakker & Wicherts, 2011; Ioannidis, 2005; Pashler & Wagenmakers, 2012).

This work must be considered within regard to concerns about the quality of research reports. *Inconsistencies* (or *congruency errors*) occur when the reported *p*-value does not correspond to the reported test statistic and its degrees of freedom. Bakker and Wicherts (2011) found that inconsistencies were quite common in international journals of psychology: 18% of the reported statistical results were inconsistent; and in approximately 1 out of 7 papers reviewed, at least one conclusion appears to have been unfounded on the basis of the results presented. Similar results have been found in Spanish psychological journals (Caperos & Pardo, 2013) and in medical journals (Berle & Starcevic, 2007; García-Berthou & Alcaraz, 2004). More importantly, some of the inconsistencies detected require a change in at least one conclusion after recalculating the *p*-value: between 3% and 23% of papers (this result varies with the study) include some of these gross inconsistencies (Berle & Starcevic, 2007; Caperos & Pardo, 2013). Furthermore, it should be noted that these inconsistencies might not only affect the particular studies' conclusions, but also the outcome of meta-analytic research that includes this

information (Bakker & Wicherts, 2011; Caperos & Pardo, 2013).

While several studies have evaluated the prevalence of inconsistencies related to the Student-*t*, Fisher-*F* of ANOVA, and Pearson's Chi-Square tests (Bakker & Wicherts, 2011; Berle & Starcevic, 2007; Caperos & Pardo, 2013; García-Berthou & Alcaraz, 2004), no studies have been carried out which evaluate the prevalence of inconsistencies in correlation analysis reports.

Pearson's product-moment correlation coefficient and others, such as Spearman's rho, indicate the magnitude and direction of the linear relationship between two quantitative variables. The null hypothesis usually tested in correlation analysis is that the value of the coefficient is zero in the population. While the application of this test is widespread in the social and health sciences, its use is not problem free. Porter (1999) and Onwuegbuzie and Daniel (2002) reported several kinds of frequent errors in the use of correlation analysis, such as the inadequate proof of statistical assumptions, or inferring causation from correlation coefficients. The main objective of this paper is to provide evidence of another kind of error, that is, the inconsistencies in the reported *p*-values. More specifically, our objective is to obtain evidence of the prevalence of inconsistencies in correlation analysis and their impact on the meta-analysis that incorporates them.

We also propose to obtain an estimate of how researchers behave regarding two topics related to the NHST strategy in general, and with the correlation analysis in particular: (1) the use of some class of control over the Type I error rate when carrying out multiple comparisons and (2) the decision about whether to perform a one- or two-tailed test.

Firstly, many sources, including prestigious handbooks of statistics and data analysis (e.g., Hays, 1994; Howell, 2002; Keppel & Wickens, 2004; Kirk, 2013; Maxwell & Delaney, 2004; Winer, Brown, & Michels, 1991), recommend that some class of adjustment be applied to control the Type I error rate when simultaneous inference is carried out. In the case of correlation tests, the *p*-value can be adjusted via Bonferroni or Sidák inequalities (Cabin & Mitchell, 2000; Curtin & Schulz, 1998) or other more powerful strategies (Hochberg & Tamhane, 1987; Holm, 1979; Keselman, Cribbie, & Holland, 1999; Keselman, Miller, & Holland, 2011; Rousselet & Pernet, 2012). The argument for applying these adjustments is that the Type I error rate increases as the number of tests increases (Curtin & Schulz, 1998; Rice, 1989; Wright, 1992). Nevertheless, there is no agreement about the ideal way of adjusting the Type I error rate (see, e.g., Keselman et al., 2011, p. 421), nor is there agreement about whether some kind of control should be applied (Rothman, 1990). However, regardless of whether or not some control should be applied, the focus of this study is to gather information about the decision made by researchers to use some control, and how the results obtained are changed by its use.

Secondly, because the effect tested in a correlation analysis is directional, that is, the relationship may be positive or negative, a two-tailed test could be replaced by a one-tailed test, if the researcher is only interested in one kind of relationship. Most statistics and data analysis books recommend using one-tailed tests when certain conditions are met. For example, Winer et al. (1991) recommend using one-tailed tests "in cases in which the experimenter is interested in rejecting H_0 only when the alternative hypothesis is one having a specified direction" (p. 44), and Zar (2004) points out that a one-tailed test is appropriate if there is a "reason *a priori* to hypothesize that a change... would be in one specified direction" (p. 579). However, there is no widespread agreement about this (see Lombardi & Hurlbert, 2009). Our intention with regard to this issue is not to denounce bad practices, but to collect information about how researchers behave and how their behavior affects the reported *p*-values.

The first aim of this work is to estimate the frequency of correlation-analysis-related inconsistencies in four Spanish journals of psychology indexed in the Journal Citation Reports (JCR). This aim involves:

- evaluating the characteristics of the reports that include correlation analysis (specifically if the information provided includes statistics with their degrees of freedom and exact *p*-values);
- checking the consistency of the reported statistical results, which is to say, the existing congruence between the reported *p*-value and the value of the test statistic and its degrees of freedom;
- evaluating whether the inconsistencies detected affect the conclusions of the study; and
- estimating how these inconsistencies can affect the meta-analysis that incorporates them.

Additionally, we are interested in collecting information about the behavior of researchers in relation to two topics:

- the use of some class of adjustment to control the Type I error rate when carrying out multiple comparisons; and
- the use of one-tailed tests in the presence of a directional hypothesis.

Method

Sample

From the Spanish journals of psychology indexed in the 2009 *Journal Citations Report* (*Social Science Edition*), we

Table 1. Journals, volumes, and papers checked, and number of statistics compiled per volume

Journal	Volume	Number of papers	Number of valid papers	Observed statistics
An Psicol	27 (1)	30	11	427
	27 (2)	32	10	476
Psicológica	32 (1)	7	1	16
	32 (2)	13	2	19
	33 (1)	7	1	15
Psicothema	23 (1)	25	8	296
	23 (2)	26	5	135
Span J Psychol	14 (1)	46	15	566
Total		186	53	1,950

selected those with a more general or multidisciplinary aim: *Anales de Psicología*, *Psicológica*, *Psicothema*, and the *Spanish Journal of Psychology*. We reviewed all the papers in each journal in every volume published in 2011. As a consequence of the lower number of papers per volume in the journal *Psicológica*, we also included one volume published in 2012. However, more papers were reviewed from some journals than from others. Table 1 shows the specific data of the reviewed volumes and the number of papers per volume. From the 186 reviewed articles we recorded 1,950 coefficients of correlation distributed in 53 papers (Mean = 36.8, *SD* = 38.0, min = 2, max = 204).

Compilation of Information

We included in the study all those papers reporting at least one correlation test. We excluded from the study those correlation analyses in which statistical significance was not the relevant question, such as in correlation analyses previous to exploratory or confirmatory factorial analysis, or correlations between subscales of the same test. We recorded the following information from each paper:

- *Reported details of correlation test*. From each reported correlation test we recorded the correlation coefficient value (*r*), the sample size, and the reported *p*-value (*complete* report). If the sample size was not available we tried to obtain it from the descriptions in the method section (*incomplete* report).
- *Use of one- versus two-tailed test*. We have recorded the presence of explicit affirmations about the application of one- or two-sided tests. When such affirmations were not included, we have assumed that the authors were implementing a bilateral test. In order to evaluate possible inconsistencies regarding this topic, we checked whether the authors make explicit directional predictions about relationships between variables, that is, if the authors express interest in whether the relationship is positive or negative, but not both (we also considered a prediction as directional in the case of evaluating the convergent validity of instruments and in the testing of specific models).
- *Control of Type I error rate*. Finally, we recorded whether some class of adjustment has been implemented to control the Type I error rate when carrying out multiple comparisons. We have taken a related group of observations whose statistical analyses are within a mathematical framework to be correlations within the same family (Miller, 1981, p. 34). Therefore, we used the referent most recommended by experts, namely, the *familywise error rate* (see, e.g., Keppel & Wickens, 2004, pp. 112–113; Kirk, 2013, p. 162), to collect this information. By way of explanation several examples are provided below. In Pedrero et al. (2011) we considered the existence of two families of correlations: the first addressing the relationships between TCI-R-67 and FrSBe-Sp scales (Table 2 in the study); the second addressing the relationship between personality traits and addiction variables (Table 3 in the study). In Delgado, Oliva, and Sánchez-Queija (2011) we considered that the correlations presented in Tables 1 and 2 belong to the same family because they answer related questions, with the same sample and the same statistical tests. Finally, in Rodríguez-Biglieri and Vetere (2011, Table 4) we considered three different families of tests because observations from three different sample groups exist, that is, generalized anxiety disorder, anxious control, and control.

Prevalence of Inconsistencies

In all the recorded correlation tests from the complete and incomplete reports we recalculated the *p*-value derived from the available information, that is, from the reported sample size and *r*-value (or their transformation in *t*-value). Additionally, we recalculated the *p*-values considering the direction of research hypothesis and the number of tests performed within a family.

The recalculated values have enabled us to distinguish between: (a) the inconsistencies based on the originally reported *p*-values; (b) the inconsistencies based on *p*-values that would have been obtained if a unilateral contrast (when this test would have been appropriate) had been performed; and (c) the inconsistencies based on *p*-values that would have been obtained by applying the Bonferroni strategy to correct the error rate per family of comparisons.

After comparing original and recalculated data, we classified each correlation test into one of the following categories (Caperos & Pardo, 2013):

- *No inconsistency*: the reported result coincides with our calculations based on the available information;
- *Slight inconsistency*: the inconsistency detected did not lead to a change in the conclusion (e.g., using $p = .232$ instead of $p = .198$, or $p = .002$ instead of $p = .007$); and
- *Gross inconsistency*: the inconsistency detected alters the statistical conclusion and changes rejection into non-rejection, or non-rejection into rejection (e.g., using $p = .14$ instead of $p = .014$, or $p < .05$ instead of $p = .086$).

Consequences of Inconsistencies on Meta-Analysis

In order to evaluate how the inconsistencies may alter the results of a meta-analysis, we compared, in all cases where inconsistencies were detected, the reported *r*-value (i.e., the observed effect size) with the *r*-value corresponding to a sample size and a *p*-value as the reported (i.e., estimated effect size). For example, if $n = 50$ and $p = .01$, then $t = 2.11$ and $r = .29$, therefore the estimated effect size is .29. In the case of inconsistencies based on inexact *p*-values, we estimated the effect size by using the *r*-value associated with the reported level of significance. For example, if $n = 30$, $r = .445$, and $p < .01$, then the estimated *r*-value is .463, that is, the *r*-value corresponding to $p = .01$.

To rate the discrepancies between the observed and recalculated *r*-values, we used the cut points proposed by Cohen (1992) to identify *small*, *medium*, and *large* effect sizes, that is, .1, .3, and .5, respectively.

Inter-Rater Agreement

The first author compiled the information. The second and third authors reviewed the papers for inclusion or exclusion in the study. This was followed by a review of directionality of the research hypothesis and the classifications in family test of the selected papers. All three authors discussed discrepancies until 100% agreement was reached (inclusion-exclusion: one discrepancy; hypothesis directionality: six discrepancies; family test: five discrepancies). In addition, the second author also reviewed 30% of the correlations included in the study reaching 98% agreement on data compilation. Finally, the three authors reviewed all the inconsistencies detected in order to ascertain that they had originated from the original paper.

Results

Reported Details of Correlation Test

A total of 1,950 correlation tests were analyzed: 1,852 (95.0%) based on Pearson coefficients, 38 (1.9%) based on Spearman coefficients, and 60 (3.1%) based on partial coefficients. Of these tests, the report was *complete* (i.e., the sample size and *p*-value had been included) in 283 (14.5%) and *incomplete* (i.e., the sample size was missing) in 1,634 (83.8%), although in these tests the sample size could be ascertained from the information included in the method. We classified 33 (1.7%) tests as *nonvalid* because the authors had not included the sample size (12 tests), the *p*-value (15 tests), or the exact *r*-value (6 tests) in the report. This resulted in the elimination of 33 tests from the prevalence of inconsistencies study because the papers did not contain the information required.

Most of the tests, 1,790 (92.5%), included an inexact *p*-value, while only 138 (7.2%) tests reported an exact *p*-value (32 tests) or the minor *p*-value recommended by APA (i.e., $p < .001$; APA, 2010, p. 114; 106 tests). Finally, 7 (0.4%) *p*-values were *implausible* ($p = 0$).

Use of One- Versus Two-Tailed Test

The authors expected a positive or negative relationship between variables (directional hypotheses) in 1,331 (68.3%) of the 1,950 valuated tests. However, none of the authors had indicated that they had used a one-tailed test.

Control of Type I Error Rate

With one exception, all the correlation tests were presented within a family of tests. Papers included 1–4 families of tests (Mean = 1.4, $SD = 0.7$). The mean number of tests per family was 25.7 ± 33.8 tests, the smallest family consisted of two tests, and the largest family of 204 tests. Only in 74 (3.8%) tests (from two different papers) did the authors control for Type I errors, using the *Bonferroni* procedure in all cases. In the 53 papers including some correlation test, we detected 76 families of tests.

Prevalence of Inconsistencies

We obtained 1,917 valid correlation tests (i.e., those tests from which we could compile the sample size, *r*-value, and *p*-value). Of these, the null hypothesis was rejected in 1,332 (69.5%) tests. The information presented was consistent in 1,813 (94.6%) tests; 27 (1.4%) tests included a *slight inconsistency*; and 77 (4.0%) tests included a *gross*

inconsistency (i.e., an inconsistency whose correction implied a change in the conclusion). Of these 77 gross inconsistencies, 35 caused a change from non-rejection to rejection and 42 caused a change from rejection to a non-rejection. As a result of this, the recalculation of inconsistent tests led to 1,325 (69.1%) rejections. Gross inconsistencies appear in 18 articles (mean per article 4.3 ± 3.8), two of which contained 11 inconsistencies each.

Given that most of the *p*-values reported were inexact, the prevalence of possible inconsistencies due to the use of two-tailed tests instead of a one-tailed test in the presence of a directional hypothesis, and due to the absence of control over the Type I error rate was evaluated using the exact *p*-values recalculated from the reported *r*-value and sample size. For this analysis we used the 1,840 correlation tests which were free from gross inconsistencies. Firstly, had the authors taken into account the direction of the hypothesis tested, they would have rejected the null hypothesis instead of maintaining it in 37 (2.0%) tests. Secondly, when applying a conservative method (*Bonferroni*) to protect against Type I errors due to multiple tests, 293 (15.9%) of the rejected hypotheses should have been maintained. These 293 inconsistencies are distributed in 40 test families. In 9 of these families more than 10 inconsistencies are observed and 73 (40 + 33) of the 293 inconsistencies appear in only two families (of 118 and 91 tests). Nonetheless, by applying a less conservative approach than that of Bonferroni, for example by using Bonferroni with families with five tests or less, and α .01 instead of .05 with families with more than five tests, the 293 inconsistencies would be reduced to 128 (7.0%).

Consequences of Inconsistencies on Meta-Analysis

In the 104 inconsistencies detected (27 slight plus 77 gross) we calculated the expected *r*-value associated with a sample size and *p*-value as reported. The mean difference between the observed and expected *r*-values was .09 (±.10), with .39 being the largest difference. From the 104 inconsistencies, 76 (73.1%) can be considered to be insignificant (with discrepancies < .1), 21 (20.2%) can be considered to be relevant (with discrepancies between .1 and .3), with 7 (6.7%) greater than .3.

Discussion

The information collected in this study about correlation tests indicates that statistical inconsistencies in peer-reviewed journals do not appear to be infrequent.

Correlation coefficient tests appeared in 28.5% of the papers reviewed and are present in a large number per paper (36.8 ± 38.0). Nevertheless, despite the high frequency of use of these tests, 4.0% of statistical conclusions could be erroneous. Taking into account all the types of inconsistencies considered in our study, approximately 13.0% of statistical conclusions could be erroneous: 77/1,917 (4.0%) due to gross inconsistencies; 37/1,840 (2.0%) due to inconsistencies related to directionality of test; and 128/1,840 (7.0%) due to inconsistencies related to the control of the Type I error rate.

The prevalence of inconsistencies in reported *p*-values has been described in several papers for the *t*-test, *F*-ANOVA, and chi-square statistics (Bakker & Wicherts, 2011; Berle & Starcevic, 2007; Caperos & Pardo, 2013; García-Berthou & Alcaraz, 2004). Inconsistencies in low impact factor journals, that is, those with an impact factor below 1.5, appear in 10.3%–21.3% of reported hypothesis tests (Bakker & Wicherts, 2011). In this study we found that inconsistencies when reporting correlations appear less frequently (5.4%) than in reports concerning other tests; but it is important to bear in mind that because correlation reports are nearly always presented with an inexact *p*-value, the majority of *slight* inconsistencies cannot be detected (Bakker & Wicherts, 2011; Caperos & Pardo, 2013). With regard to inexact *p*-values, the following prevalence of inconsistencies was found: 4.7% (Caperos & Pardo, 2013), and 7.5%–8.1% (Bakker & Wicherts, 2011), which are very similar values to those found in our study on correlation tests.

When considering only gross inconsistencies (inconsistencies that imply a change in the statistical conclusion), we found that these accounted for 4%. The prevalence of gross inconsistencies is similar between the exact and inexact *p*-values, and has been found to appear in 3% of the tests (Caperos & Pardo, 2013). Bakker and Wicherts (2011) found a mean value of gross inconsistencies between 1.6% and 3.4% in low impact journals. Our data also shows a prevalence of gross inconsistencies that is similar to those found in other statistical tests.

Regarding the use of the one-tailed test, we found that two-tailed tests were more frequently used even when authors were testing directional hypotheses. Good and Hardin (2006) considered that confusion over the use of one- versus two-tailed tests is a common error in statistics; and we cannot ignore that the use of one-tailed tests is controversial (Kimmel, 1957; Lombardi & Hurlbert, 2009; Rice & Gaines, 1994; Ruxton & Neuhäuser, 2010). Several authors propose a restrictive use of one-tailed tests considering that they are justified only when it "is absolutely certain that a result in the unpredicted direction is impossible" (Lombardi & Hurlbert, 2009). Regardless of this concern, our study shows that 2.0% of null hypotheses in the

reviewed journals would change if the authors had chosen a one-tailed test.

Papers reporting correlations usually include several tests between different variables measured in the same sample. We found between one and four different families of tests per paper. Authors from two different papers controlled for Type I errors only in 74 (3.8%) tests. In these cases the authors considered the correlations between one variable and several others as belonging within the same family, but did not consider as a family the complete set of correlations within the same sample. There is some controversy about what can be considered a family of tests, and subjective points of view can intervene in the decision (Miller, 1981). In our paper we have adopted Miller's (1981, p. 34) proposal: "the natural family (...), in the majority of instances, is the individual experiment of a single researcher." Regardless of the more or less strict definition of family of tests, it is important for researchers to evaluate the occurrence of Type I errors when many tests are performed using the same data. The implementation of the *Bonferroni* correction in our sample of reviewed papers would lead to maintaining 15.9% of the rejected hypotheses. A less conservative strategy (α = .01) would lead to maintaining 7.0% of the rejected hypotheses. Olejnik, Li, Supattathum, and Huberty (1997) affirm that differences regarding the number of hypotheses rejected with real data sets among different procedures to adjust for Type I errors are less than 4%. The aim of this work is not to review all the procedures designed to control a Type I error rate, but to inform authors and editors about their use (see Olejnik et al., 1997, for classical methods; or Keselman et al., 2011, for new approximations). Several authors have recently highlighted the importance of establishing adequate practices to reduce false discoveries (Ioannidis, 2005, 2008; Pashler, & Wagenmakers, 2012; Wagenmakers, Wetzels, Borsboom, & Van Der Maas, 2011). It is the responsibility of authors to improve the review process when publishing results, in order to reduce the prevalence of inconsistencies. But it is also the responsibility of editorial boards to adopt clear policies concerning how to proceed when working with multiple correlation tests and, in general, with multiple inferences.

The fact that, on many occasions, correlation analysis is used as an exploratory strategy may reduce the importance of inconsistencies related to statistical significance. However, meta-analytic reviews compile effect sizes regardless of their level of significance (Botella & Sánchez-Meca, 2015; Card, 2011). A common practice in meta-analytic reviews is to collect effect sizes (*r*-values) from large tables reported in published papers that include the relationship studied. As a result of an inconsistency, an erroneous *r*-value could be included in a meta-analysis. We found a mean discrepancy of .09 ± .1 point between the expected and observed *r*-values (excluding in this the inconsistencies related to the directionality of test and those related to the control of the Type I error rate). Moreover, 26.9% of the inconsistencies might lead to the inclusion in meta-analysis of moderate or strong discrepancies in effect sizes. These results are similar to those of Bakker and Wicherts (2011), who found that discrepancies in Cohen's *d* owing to inconsistencies in *t*-tests and two groups of ANOVA comparisons would lead to important errors in 23% of cases.

Based on the sample of papers included in this study, we can conclude that statistical reports concerning correlation tests can be improved upon. With reference to the objectives of this study we found:

- most of the correlation reports did not include the sample size (83.8%) or the exact *p*-value (92.5%);
- gross inconsistencies appear at a similar rate (4%) to that previously found for *t*-tests, chi-square tests, or *F*-tests; and
- 26.9% of inconsistencies could bias the meta-analysis that includes them. In addition, the use of one-tailed tests is negligible, even when the researcher's hypothesis is directional; and a large number of Type I errors (16.6%) due to multiple inferences without any control in the error rate were found.

Authors, in particular, but also reviewers and editorial boards are encouraged to pay particular attention to prevent inconsistencies in statistical reports in Spanish psychology journals.

References

APA. (2010). *Publication manual of the American Psychological Association*, (6th ed.). Washington, DC: Author.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavioral Research Methods, 43*, 666–678. doi: 10.3758/s13428-011-0089-5

Bakker, M., & Wicherts, J. M. (2014). Outlier removal and the relation with reporting errors and quality of psychological research. *PLoS One, 9*, e103360.

Balluerka, N., Gómez, J., & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*, 55–70. doi: 10.1027/1614-1881.1.2.55

Berle, D., & Starcevic, V. (2007). Inconsistencies between reported test statistics and *p*-values in two psychiatry journals. *International Journal of Methods in Psychiatric Research, 16*, 202–207. doi: 10.1002/mpr.225

Botella, J., & Sánchez-Meca, J. (2015). *Qué es el meta-análisis* [What is the meta-analysis?]. Madrid, Spain: Síntesis.

Cabin, R. J., & Mitchell, R. J. (2000). To Bonferroni or not to Bonferroni: When and how are the questions. *ESA Bulletin, 81*, 246–248.

Caperos, J. M., & Pardo, A. (2013). Consistency errors in *p*-values reported in Spanish psychology journals. *Psicothema, 25*, 408–414. doi: 0.7334/psicothema2012.207

Card, N. A. (2011). *Applied meta-analysis for social science research*. New York, NY: Guilford Press.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. doi: 10-1037/0033-2909.112.1.155

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., … Wilson, S. (2007). Statistical reform in psychology. Is anything changing? *Psychological Science, 18*, 230–232. doi: 10.1111/j.1467-9280.2007.01881.x

Curran-Everett, D. (2000). Multiple comparisons: Philosophies and illustrations. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology, 279*, 1–8.

Curtin, F., & Schulz, P. (1998). Multiple correlations and Bonferroni's correction. *Biological Psychiatry, 44*, 775–777. doi: 10.1016/S0006-323(98)00043-2

Delgado, I., Oliva, A., & Sánchez-Queija, I. (2011). Apego a los iguales durante la adolescencia y la adultez emergente [Attachment to peers in adolescence and emerging adulthood]. *Anales de Psicología, 27*, 155–163.

García-Berthou, E., & Alcaraz, C. (2004). Incongruence between test statistics and *p*-values in medical papers. *BMC Medical Research Methodology, 4*, 13–17. doi: 10.1186/1471-2288-4-13

Good, P. I., & Hardin, J. W. (2006). *Common errors in statistics (and how to avoid them)*. Hoboken, NJ: Wiley.

Hays, W. L. (1994). *Statistics* (5th ed.). New York, NY: Holt, Rinehart and Winston.

Hochberg, Y., & Tamhane, A. (1987). *Multiple comparison procedures*. New York, NY: Wiley.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal Statistics, 6*, 65–70.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Belmont, CA: Thomson Wadsworth.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124. doi: 10.1371/journal.pmed.0020124

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*, 640–648. doi: 10.1097/EDE.0b013e31818131e7

Jeličić, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology, 45*, 1195–1199. doi: 10.1037/a0015665

Keppel, G., & Wickens, Th. D. (2004). *Design and analysis. A researcher's handbook* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Keselman, H. J., Cribbie, R., & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise/comparisonwise Type I error control. *Psychological Methods, 4*, 58–69. doi: 10.1037/1082-989X.4.1.58

Keselman, H. J., Miller, C. W., & Holland, B. (2011). Many tests of significance: New methods for controlling Type I errors. *Psychological Methods, 16*, 420–431. doi: 10.1037/a0025810

Kimmel, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin, 54*, 351–353.

Kirk, R. E. (2013). *Experimental design. Procedures for the behavioral sciences* (4th ed.). Thousand Oaks, CA: Sage.

Lombardi, C. M., & Hurlbert, S. H. (2009). Misprescription and misuse of one-tailed tests. *Austral Ecology, 34*, 447–468.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data* (2nd ed.). Mahwah, NJ: LEA.

Miller, R. G. (1981). *Simultaneous statistical inference*. New York, NY: McGraw Hill.

Nikerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241–301.

Olejnik, S., Li, J., Supattathum, S., & Huberty, C. J. (1997). Multiple testing and statistical power with modified Bonferroni procedures. *Journal of Educational and Behavioral Statistic, 22*, 389–406. doi: 10.3102/10769986022004389

Onwuegbuzie, A. J., & Daniel, L. G. (2002). Uses and misuses of the correlation coefficient. *Research in the Schools, 9*, 73–90.

Pardo, A., Garrido, J., Ruiz, M. A., & San Martín, R. (2007). La interacción entre factores en el análisis de varianza: errores de interpretación [The interacction between factors in ANOVA: Misconceptions]. *Psicothema, 19*, 343–349.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*, 528–530. doi: 10.1177/1745691612465253

Pedrero, E. J., Ruiz, J. M., Olivar, Á., Rojo, G., Llanero, M., & Puerta, C. (2011). Diferencias de personalidad entre adictos al alcohol y controles emparejados: relación con sintomatología frontal y subtipos de adictos [Personality differences between alcohol adicts and matched controls: Relationship with frontal symptoms and subtypes of addicts]. *Psicothema, 23*, 100–106.

Porter, A. M. (1999). Misuse of correlation and regression in three medical journals. *Journal of the Royal Society of Medicine, 92*, 123–128. doi: 10.1177/ 014107689909200306

Rice, W. R. (1989). Analyzing tables of statistical tests. *Evolution, 43*, 223–225.

Rice, W. R., & Gaines, S. D. (1994). "Heads I win, tails you lose": Testing directional alternative hypotheses in ecological and evolutionary research. *Trends in Ecology & Evolution, 9*, 235–237. doi: 10.1016/0169-5347(94)90258-5

Rodríguez-Biglieri, R., & Vetere, G. L. (2011). Psychometric characteristics of the Penn State Worry Questionnaire in an Argentinean sample: A cross-cultural contribution. *The Spanish Journal of Psychology, 14*, 452–463. doi: 10.5209/rev_SJOP.2011.v14.n1.41

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *The American Psychologist, 44*, 1276–1284. doi: 10.1037/0003-066X.44.10.1276

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology, 1*, 43–46.

Rousselet, G. A., & Pernet, C. R. (2012). Improving standards in brain-behavior correlation analyses. *Frontiers in Human Neuroscience, 6*, 231–243. doi: 10.3389/fnhum.2012.00119

Ruxton, G. D., & Neuhäuser, M. (2010). When should we use one-tailed hypothesis testing? *Methods in Ecology and Evolution, 1*, 114–117. doi: 10.1111/j.2041-210X.2010.00014.x

Veldkamp, C. L., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PLoS One, 9*, e114876. doi: 10.1371/journal.pone.0114876

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & Van Der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*, 426–432. doi: 10.1037/a0022790

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York, NY: McGraw-Hill.

Wright, S. P. (1992). Adjusted *p*-values for simultaneous inference. *Biometrics, 48*, 1005–1013. doi: 10.2307/2532694

Zar, J. H. (2004). *Biostatistical analysis* (5th ed.). New York, NY: Prentice-Hall.

Received March 10, 2014
Revision received May 11, 2015
Accepted December 16, 2015
Published online June 20, 2016

Jose Manuel Caperos is an associated professor at the Centro San Rafael-Nebrija (Madrid, Spain) teaching statistics and research methodology. His current research interests include analysis of common statistical and methodological mistakes in psychology and health sciences research, and the study of human behavior from an evolutionary perspective.

Ricardo Olmos is an associated professor of data analysis at the Department of Social Psychology and Methodology at the Universidad Autónoma de Madrid, Spain. His research interests lie in applied statistics, specially missing data imputation methods, and automatic essay assessment with computational linguistic models.

Antonio Pardo is a full professor of data analysis at the Department of Social Psychology and Methodology at the Universidad Autónoma de Madrid, Spain. His research is focused in applied statistics, the appropriate use of statistics, and the assessment of clinical change.

Antonio Pardo
Department of Psychology
Universidad Autónoma de Madrid
Cantoblanco
28049 Madrid
Spain
Tel. +34 91 497-4061
Fax +34 91 497-6211
E-mail antonio.pardo@uam.es

