Psicothema

Psicothema 2013, Vol. 25, No. 3, 408-414
doi: 10.7334/psicothema2012.207

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Consistency errors in p-values reported in Spanish psychology journals

José Manuel Caperos[1] and Antonio Pardo[2]
[1] Universidad Pontificia de Comillas and [2] Universidad Autónoma de Madrid

Abstract

Background: Recent reviews have drawn attention to frequent consistency errors when reporting statistical results. **Method:** We have reviewed the statistical results reported in 186 articles published in four Spanish psychology journals. Of these articles, 102 contained at least one of the statistics selected for our study: Fisher-*F*, Student-*t* and Pearson-χ^2. **Results:** Out of the 1,212 complete statistics reviewed, 12.2% presented a consistency error, meaning that the reported p-value did not correspond to the reported value of the statistic and its degrees of freedom. In 2.3% of the cases, the correct calculation would have led to a different conclusion than the reported one. In terms of articles, 48% included at least one consistency error, and 17.6% would have to change at least one conclusion. In meta-analytical terms, with a focus on effect size, consistency errors can be considered substantial in 9.5% of the cases. **Conclusion:** These results imply a need to improve the quality and precision with which statistical results are reported in Spanish psychology journals.

Keywords: statistical errors, null hypothesis significance testing, *p*-value.

Resumen

Errores de consistencia en los valores p informados en las revistas españolas de Psicología. **Antecedentes:** recientes trabajos han llamado la atención sobre la presencia de frecuentes errores de consistencia al informar de los resultados estadísticos. **Método:** hemos revisado los resultados estadísticos de 186 artículos publicados en cuatro revistas españolas de Psicología, 102 de los cuales contenían alguno de los estadísticos seleccionados para nuestro estudio: *F* de Fisher, *t* de Student y χ^2 de Pearson. **Resultados:** de los 1.212 contrastes completos revisados el 12,2% presenta algún tipo de error de consistencia: el valor *p* informado no se corresponde con el valor del estadístico de contraste y sus grados de libertad. En el 2,3% de los casos el error detectado conllevaría un cambio en la conclusión estadística del contraste. En términos del número de artículos, el 48% de los revisados incluye algún error de consistencia y al menos el 17,6% tendría que cambiar alguna de sus conclusiones. En términos meta-analíticos, los errores de consistencia pueden considerarse importantes en el 9,5% de los casos. **Conclusiones:** estos resultados señalan la necesidad de mejorar la calidad y precisión con la que se informa de los resultados estadísticos en las revistas españolas de Psicología.

Palabras clave: errores estadísticos, contraste de hipótesis, valor *p*.

The purpose of the majority of studies published in the field of social and health sciences is to empirically test one or several hypotheses. The dominant strategy used to reach this goal is *null hypothesis significance testing* (Fisher, 1925, 1935; Neyman & Pearson, 1928). This strategy consists of maintaining or rejecting a hypothesis, referred to as *null hypothesis,* which affirms that the searched effect does not exist. The null hypothesis is maintained or rejected depending on its degree of compatibility with the empirical data, assessed in terms of probability. Despite the critique that this approach has received (see, for example, Cohen, 1994; Nikerson, 2000; Wagenmakers, 2007), the approach continues to be the most common one in data analyses reported in the field of psychology. According to Cumming et al. (2007), null hypothesis testing is used in more than 95% of the empirical articles published in psychology journals.

Irrespective of the appropriateness of this strategy to test hypotheses, specialized literature has repeatedly pointed out that it is not infrequent to find that research reports include errors related to the way of analyzing data and to the way of interpreting them (Curran-Everett, 2000; Pardo, Garrido, Ruiz, & San Martín, 2007; Rosnow & Rosenthal, 1989; Jeličić, Phelps, & Lerner, 2009). Furthermore, recent reviews in different areas of knowledge have called attention to the frequent presence of inconsistencies between reported test statistics and *p*-values (García-Berthou & Alcaraz, 2004; Berle & Starcevic, 2007; Bakker & Wicherts, 2011).

The *American Psychological Association* (APA) recommends including all data needed to assess the used statistics in the report. Among other things, it recommends informing about the value of the test statistic, the degrees of freedom, and the exact *p*-value (APA, 2010, p. 34). Consistency errors occur when the reported *p*-value does not correspond to the *p*-value associated with the reported test statistic and with its degrees of freedom (Bakker & Wicherts, 2011). These errors can be due to simple mistakes in copying or in reading the output of the used statistical program, but they can also be due to a lack of knowledge about the applied procedures. Examples of the latter case may occur when, for an ANOVA, the total degrees of freedom are reported instead of the degrees of freedom of the error, or when, for the *t* statistic, an incorrect decision is made because a two-sided *p*-value is used in a one-sided test.

Received: March 24, 2012 • Accepted: July 31, 2012
Corresponding author: José Manuel Caperos
Faculty of Humanities and Social Sciences
Comillas Pontifical University
28049 Madrid (Spain)
e-mail: jcaperos@upcomillas.es

García-Berthou and Alcaraz (2004) indicated that 11.6% of the statistical results in *Nature* and 11.1% in the *British Medical Journal* are inconsistent. These same authors reported that 38% of the reviewed articles in *Nature* and 25% of the articles in the *British Medical Journal* contain at least one consistency error. In the field of psychiatry, Berle and Starcevic (2007) observed that 14.3% of the reviewed statistical results are inconsistent, with 36% of the reviewed articles having one or more consistency errors.

In an important number of cases, consistency errors imply a change in the test conclusion. In a review of different psychology journals, Bakker and Wicherts (2011) found that 15% of the articles include at least one statistical conclusion that ought to be changed after correctly calculating the *p*-value. Bakker and Wicherts highlight the fact that among the errors that affect the statistical conclusion, the ones that go in the direction of declaring non-significant results significant are more frequent than the ones that go in the opposite direction, indicating the presence of a bias in favor of the researcher's expectations.

Consistency errors in statistical results also affect meta-analytical reviews that include the reports with the errors. Many experts recommend using effect size measures and including them in research reports, accompanying statistical significance (Abelson, 1995; APA, 2010; Cohen, 1988; Cummings et al., 2007; Murphy, 1997; Thomson, 1994, 1997). The recommendations in the report by Wilkinson and the Task Force on Statistical Inference (1999) are especially relevant. Even so, the inclusion of effect size measures in research reports is not common practice. The available reviews indicate that between 30% and 60% of the articles do not include any effect size measure (Sun, Pan, & Wang, 2010; McMillan & Foley, 2011), reaching 93% in the case of mean comparisons (Zientek, Capraro, & Capraro, 2008). When these measures are not reported, the estimations of the effect size necessary to elaborate a meta-analysis are based on the test statistic and its degrees of freedom (Botella & Gambara, 2002; Card, 2012; Sánchez-Meca & Botella, 2010). Bakker and Wicherts' (2011) results indicate that, when calculating Cohen's *d*, the difference between the value calculated with errors and the one calculated without errors exceeds 0.10 points in 23% of the evaluated cases. According to the authors, that difference can importantly affect the results of a meta-analysis.

The first aim of this work is to estimate the frequency of consistency errors in four Spanish psychology journals indexed in the *Journal Citation Reports* (JCR). This aim implies: (a) evaluating the characteristics of the reports that include statistical results (namely, if the information offered includes statistics with their degrees of freedom, exact *p*-values, and effect size measures) and (b) assessing the consistency of the reported statistical results, which is to say, the existing congruence between the value of the used statistic with its degrees of freedom and the reported *p*-value. We also offer a classification of the observed consistency errors as well as an approximation to their possible causes. Our second aim is to evaluate how consistency errors affect the conclusions of the reports that include them and the meta-analytical studies that incorporate their results.

Method

Sample

From the Spanish journals of psychology indexed in the *Journal Citation Reports* from 2009 (*Social Science Edition*), the ones with a more general or multidisciplinary aim were selected: *Anales de Psicología*, *Psicológica*, *Psicothema,* and *Spanish Journal of Psychology*. All articles published in 2011 within each journal were selected, but we added one 2012 volume from *Psicológica* due to the reduced number of articles per volume in this journal. Even so, more articles were reviewed from some journals than from others. Table 1 shows the specific data of the reviewed volumes and the number of articles per volume.

Procedure

Information compilation. We collected information from three statistical tests: the ANOVA's *F*, Student's *t*, and Pearson's χ^2. As in Bakker and Wicherts' study (2011), *F* and *t* statistics of regression analyses were not taken into consideration (because they are not always reported), nor were χ^2 statistics used in model adjustments (because the goal in this context is to maintain the null hypothesis, not to reject it).

We selected these statistics because they are the most frequently used ones in the area of psychology (Berle & Starcevic, 2007) and because they are the ones that have been used in studies similar to ours (Bakker & Wicherts, 2011). Out of the 186 reviewed articles, 105 included one of the selected statistics: 1,717 statistical tests in total. The statistic value, the reported *p*-value, and, when possible, the degrees of freedom as well as the number and size of the groups were registered for each test. We also registered whether or not an effect size measure was included.

Quality of the provided information. The 1,717 statistics were classified in three groups, depending on the provided information:

- *Complete:* the statistic value and its degrees of freedom are included.
- *Incomplete:* the data needed to calculate the *p*-value are not explicitly reported but can be deduced from the information provided (e.g., the degrees of freedom can be obtained using the sample size and the design characteristics).
- *Non-valid*: the data needed to calculate the *p*-value are not included and cannot be deduced from the information provided.

Table 1
Journals, volumes, and papers checked, and number of statistics of each type

Journal	Volume	Num. of papers	Num. of valid papers[1]	Observed statistics			
				F	t	X^2	Total
An Psicol	27 (1)	30	17	95	105	12	212
	27 (2)	32	16	179	38	13	230
Psicológica	32 (1)	7	4	46	8	–	54
	32 (2)	13	9	145	22	18	185
	33 (1)	7	4	88	80	–	168
Psicothema	23 (1)	25	13	115	116	15	246
	23 (2)	26	13	147	40	25	212
Span J Psychol	14 (1)	46	29	265	60	85	410
Total		186	105	1,080	469	168	1,717

[1] Number of papers including at least one *F*, *t*, or X^2 statistic

Classifying a statistic as *complete* did not offer doubt. Both authors reviewed the occasional doubts (14 cases) in the classification of a statistic as *incomplete* or *non-valid* until a 100% agreement was reached.

Type of p-value. The statistics were also classified depending on the type of reported *p*-value:

- *Exact:* the exact *p*-value is offered (e.g., $p = .002$ or $p = .382$) or it's indicated that the obtained *p*-value is under .001 ($p<.001$), which is the limit below which APA (2010, p. 114) recommends not to offer the exact *p*-value.
- *Inexact:* the *p*-value is reported as greater or lower than a specific pre-established significance criterion, and that criterion is greater than .001 (e.g., $p<.01$ or $p>.05$).
- *Implausible:* the reported *p*-value is erroneous due to its impossibility (e.g., $p = .000$ or $p<.000$).

Consistency errors. In all the results classified as complete or incomplete we recalculated the *p*-value using the value of the corresponding statistic and its degrees of freedom. The *Microsoft Excel* spreadsheet was used for the calculations.

We considered that there was a consistency error when the reported *p*-value did not coincide with the *p*-value obtained by our calculations based on the available information. To decide if a specific result was a consistency error, we took the corresponding statistic's decimal precision into account. For example, if an *F* statistic appeared as $F(1, 32) = 3.5$, we considered that the true value could be any value within the 3.45-3.55 range and, consequently, that the corresponding *p*-value could be any value within the .069-.072 range. We made sure that the detected inconsistencies were not due to the use of a method of correcting the error rate by multiple comparisons. After identifying the consistency errors, each result was classified in one of four groups:

- *No error:* the reported result coincides with our calculations based on the available information.
- *Slight error:* the detected error does not lead to a change in the conclusion (e.g., using $p = .232$ instead of $p = .198$, or $p = .002$ instead of $p = .007$).
- *Moderate error:* the detected error, although not leading to a change in the conclusion, involves an important change in the degree of significance attributed to the results (e.g., using $p = .060$ instead of $p = .220$, or using $p<.01$ instead of $p = .033$).
- *Gross error:* the detected error alters the conclusion, changing the rejection of the null hypothesis into a non-rejection or a non-rejection into a rejection (e.g., using $p = .14$ instead of $p = .014$, or $p<.05$ instead of $p = .086$).

The gross and moderate errors detected in a first review by the first author were reviewed again by both authors until a 100% agreement was reached.

With the intention of proposing a practical control measure, or an improvement in reports, the *gross* errors were reviewed and classified into four groups attending to their possible cause:

- *Copy*: the error could be interpreted as a copy error from the results (e.g., informing $p = .14$ instead of $p = .014$, or, in an ANOVA, informing $gl = 7$ for a dichotomous factor).
- *One-two-tailed test*: in the case of Student's *t* statistic, making a wrong decision by using a two-sided *p*-value when the correct option would have been using a one-sided *p*-value (we have not observed the opposite error).
- *Precision in the information*: using the *lower than* sign (<) when the correct thing to do would have been using the *equal* sign (=) (e.g., using $p<.05$ instead of $p = .052$).
- *Non-identifiable cause*.

Magnitude of the consistency errors. Finally, we estimated the magnitude of the consistency errors by calculating the discrepancy between the effect size obtained using the reported statistic and the effect size obtained using the reported *p*-value. To obtain these estimations, we selected, among the tests that showed consistency errors, only the ones related to the comparison of two means (independent as well as related) using the *t* statistic (63 tests in total). The selected effect size measure was the standardized difference, calculated with the formulae proposed by Hedges (see Card, 2012).

Data analysis

To compare groups on a quantitative variable (e.g., to compare the percentage of incomplete statistics per article in the four journals), we used ANOVA's *F* (with Brown-Forsythe's correction when it was not possible to assume equal variances). For the post-hoc comparisons, Tukey's test (equal population variances) and Games-Howell's test (unequal population variances) were used. The sample size made concerns about the normality assumption unnecessary.

To compare frequencies (e.g., to compare the frequency with which each one of the three selected statistics was used) and to relate categorical variables (e.g., to relate the type of statistic with the quality of the provided information), we used Pearson's X^2 statistic (in none of these analyses did we encounter problems with the size of the frequencies). Adjusted standardized residuals were used to identify significant discrepancies between the observed and the expected frequencies.

Results

Table 1 shows the number of registered statistics: 1080 *F* statistics (62.9%), 469 *t* statistics (27.3%), and 168 χ^2 statistics (9.8%): in total, 1717 statistics. The number of statistics seems to be homogenously divided over the selected journals, $\chi^2(3) = 4.32$, $p = .229$, $V = .03$, but a significant association was observed between journal and type of statistic, $\chi^2(6) = 110.96$, $p<.001$, $V = .18$. More than expected *t* statistics were registered for *An Psicol* and *Psicothema,* more *F* statistics for *Psicológica*, and more χ^2 statistics for *Span J Psychol*.

Quality of the provided information. Of the 1717 registered statistics, 1212 were classified as *complete* (70.6%), 414 as *incomplete* (24.1%), and 91 as *non-valid* (5.3%; see Table 2). A relation between type of report and journal was observed. First, the percentage of *complete* statistics per article is not the same in all journals, $F(3, 102) = 4.53$, $p = .005$, $\eta^2 = .12$; this percentage is higher in *Psicológica* than in *An Psicol* ($p<.001$), *Psicothema* ($p<.001$), and *Span J Psychol* ($p = .048$). Second, the percentage of *incomplete* statistics per article is not the same for each journal either, $F(3, 102) = 3.49$, $p = .018$, $\eta^2 = .09$. This percentage is lower in *Psicológica* than in *An Psicol* ($p < .001$), *Psicothema* ($p = .011$), and *Span J Psychol* ($p = .012$). We did not observe significant

differences among the journals in the percentage of *non-valid* statistics per article, $F(3, 102) = 2.65$, $p = .053$, $\eta^2 = .07$.

We also found that type of report is related to type of statistic, $\chi^2(4) = 207.18$, $p<.001$, $V = .25$. The percentage of *incomplete* statistics is larger than expected with the X^2 and *t* statistics. For example, it is more common to report the degrees of freedom of the *F* statistics (82.2%) than of the χ^2 (47.0%) and *t* statistics (52.2%).

Finally, 656 of the 1717 registered statistics (38.2%) are accompanied by an effect size measure, and only 417 statistics (24.3%) are complete as well as accompanied by an effect size measure (see Table 2). Including an effect size measure is more frequent (43.7%) when using the *t* statistic and less frequent (23.8%) when using the X^2 statistic, $\chi^2(2) = 20.78$, $p<.001$, $V = .11$. Only 43 (41.0%) of the 105 reviewed articles include an effect size measure.

Type of p-value. Table 3 shows the results related to the type of *p*-value. The 825 cases (48.0%) categorized as *exact p*-values include *p*-values given with an equal sign (479; 27.9%) as well as the ones described as being below the .001 limit recommended by APA (346; 20.2%).

Type of reported *p*-value is related to journal. First, the percentage of times that an *exact p*-value is reported per article is not the same in the journals, $F(3, 102) = 3.82$, $p = .012$, $\eta^2 = .10$; this percentage is lower in *Psicológica* than in *An Psicol* ($p = .035$) and *Psicothema* ($p = .012$). Second, the percentage of times that an *inexact p*-value is reported per article is not the same in the journals, $F(3, 102) = 5.42$, $p = .002$, $\eta^2 = .14$; this percentage is higher for *Psicológica* than for *An Psicol* ($p = .033$) and *Psicothema* ($p = .01$). We did not observe differences among the journals in the percentage of times an *implausible p*-value was reported per article, $F(3, 102) = .37$, $p = .772$, $\eta^2 = .01$.

The type of the reported *p*-value is also related to the type of statistic, $\chi^2(4) = 17.41$, $p = .002$, $V = .07$; the percentage of *exact p*-values is higher than expected with the X^2 statistic and the percentage of *inexact p*-values is higher with the *F* statistic. The type of *p*-value is also related to the quality of the provided information, $\chi^2(3) = 15.42$, $p = .004$, $V = .07$; in particular, the percentage of *implausible p*-values is higher than expected among *incomplete* statistics.

Consistency errors. In our analysis of the consistency errors we did not consider the 91 statistics classified as *non-valid*. Therefore, this part of the study concerns 1626 statistics: 1,212 from *complete* reports and 414 from *incomplete* reports.

After recalculating the *p*-value corresponding to each statistic, we registered 247 errors (15.2% of the statistics). Table 4 offers details for each type of error, distinguishing *complete* and *incomplete* reports. The percentages of consistency errors are higher in *incomplete* reports than in *complete* ones, $\chi^2(3) = 38.92$, $p<.001$, $V = .15$.

Table 5 shows, for each journal, the percentage of each type of error per article. We did not observe differences among the journals in the percentages of *slight* errors, $F(3, 99) = 1.46$, $p = .230$, $\eta^2 = .04$; *moderate* errors, $F(3, 99) = 1.21$, $p = .310$, $\eta^2 = .03$; or *gross* errors, $F(3, 99) = .41$, $p = .744$, $\eta^2 = .012$. This pattern remains when considering only the *complete* statistics.

In contrast, the percentage of consistency errors is related to the type of statistic, $\chi^2(6) = 39.04$, $p<.001$, $V = .11$; the percentage of *slight* errors is larger with the χ^2 statistic and the percentage of *moderate* errors is larger with the *t* statistic. Consistency errors are also related to the type of the reported *p*-value, $\chi^2(6) = 150.59$, $p<.001$, $V = .22$; *slight* and *moderate* consistency errors are more frequent when *exact p*-values are reported than when *inexact p*-values are. This pattern remains in the *incomplete* reports; but the difference involving *moderate* errors disappears in the *complete* reports.

Table 2
Quality of the information provided

		Report			
		Complete n = 1,212	**Incomplete n = 414**	**Non-valid n = 91**	**Total n = 1,717**
Journal [1]	*An Psicol*	57.5% (43.1)	36.4% (40.7)	6.1% (19.7)	–
	Psicológica	95.3% (10.9)	1.3% (5.4)	3.4% (10.0)	–
	Psicothema	57.3% (44.9)	28.6% (40.5)	14.2% (30.5)	–
	Span J Psychol	74.8% (38.8)	25.1% (38.7)	0.1% (0.4)	–
Statistics [2]	*F*	888 (82.2%)	153 (14.2%)	39 (3.6%)	1,080
	t	245 (52.2%)	178 (38%)	46 (9.8%)	469
	X^2	79 (47%)	83 (49.4%)	6 (3.6%)	168
Effect size [2]	*Yes*	417(63.6%)	218 (33.2%)	21 (3.2%)	656
	No	795 (74.9%)	196 (18.5%)	70 (6.6%)	1,061

[1] Mean (standard deviation) of the percentage of complete, incomplete and non-valid statistics per paper
[2] Frequencies (row percentages) of each type of statistic

Table 3
Type of reported *p*-value

		Type of p-value			
		Exact n = 825	**Inexact n = 719**	**Implausible n = 172**	**Total n = 1,717**
Journal [1]	*An Psicol*	61.8% (33.8)	25.9% (31.6)	12.2% (22.1)	–
	Psicológica	33.2% (36.6)	58.6% (40.1)	8.2% (16.7)	–
	Psicothema	67.4% (29.1)	20.3% (27.3)	12.3% (23.8)	–
	Span J Psychol	51.3% (39.6)	41.0% (38.0)	7.7% (19.9)	–
Statistics [2]	*F*	497 (46.0%)	481 (44.5%)	102 (9.4%)	1,080
	t	225 (48.0%)	189 (40.3%)	55 (11.7%)	469
	X^2	103 (61.3%)	49 (29.2%)	16 (9.5%)	168
Report [2]	*Complete*	578 (47.7%)	523 (43.2%)	111 (9.2%)	1,212
	Incomplete	194 (46.9%)	162 (39.1%)	58 (14%)	414
	Not valid	53 (58.9%)	34 (37.8%)	4 (4.4%)	91

[1] Mean (standard deviation) of the percentage of exact, inexact and implausible *p*-values per paper
[2] Frequencies (row percentages) of each type of *p*-value

Table 4
Consistency errors by report type (frequencies and row percentages)

	Type of consistency error				Total	Number of errors
Report	**No error**	**Slight**	**Moderate**	**Gross**		
Complete	1064 (87.8%)	107 (8.8%)	13 (1.1%)	28 (2.3%)	1212	148
Incomplete	315 (76.1%)	63 (15.2%)	18 (4.3%)	18 (4.3%)	414	99
Total	1379 (84.8%)	170 (10.5%)	31 (1.9%)	46 (2.8%)	1626	247

Out of the 46 *gross* errors that were identified, for 22 (47.8%; 13 of them in *complete* reports) the researcher rejected the null hypothesis when the correct option would have been to maintain it, and for 24 (52.1%; 15 of them in *complete* reports) the opposite occurred. Table 6 shows a summary of the *gross* errors and their possible causes.

When focusing our attention on the articles (186) instead of on the statistics (1626), we found the following results. Of the 186 articles, 102 (54.8%) include at least one valid statistic (*complete* or *incomplete*). Of these 102 articles, 64 (62.7%) contain consistency errors: 30 articles (29.4%) contain at least one *slight* error, 10 (9.8%) contain at least one *moderate* error, and 24 (23.5%) contain at least one *gross* error. Considering only the articles with *complete* reports, 23 articles (22.5%) contain at least one *slight* error, 8 (7.8%) contain at least one *moderate* error, and 18 (17.6%) contain at least one *gross* error. It is worth mentioning that one of the considered articles contains five *gross* errors (10.9% of the total number of *gross* errors).

Consistency error magnitude. The mean effect size, as computed with the value of the statistics reported in the articles, is 0.27 (±0.24). The mean that we obtained using estimations based on the reported *p*-values is 0.34 (±0.43). In 57 of the 63 evaluated tests (90.5%) we observed small differences among the corresponding estimations of the effect size (less than 0.10 points). In the remaining 6 tests (9.5%) the differences were more substantial than 0.10 points. The largest observed difference was two points.

Table 5
Consistency errors

		Type of consistency error				
		No error	**Slight**	**Moderate**	**Gross**	**Total**
Journal[1]	*An Psicol*	83.7% (21.9)	11.8% (21.1)	1.2% (3.9)	3.3% (7.2)	–
	Psicológica	94.2% (7.0)	3.6% (5.8)	0.0% (0.2)	2.1% (3.8)	–
	Psicothema	79.1% (23.2)	15.0% (16.7)	2.5% (5.3)	3.4% (7.9)	–
	Span J Psychol	83.9% (22.5)	10.2% (18.2)	4.0% (12.7)	1.8% (4.8)	–
Statistic[2]	*F*	905 (86.9%)	95 (9.1%)	13 (1.2%)	28 (2.7%)	1,041
	t	350 (82.7%)	40 (9.5%)	16 (3.8%)	17 (4.0%)	423
	X^2	124 (76.5%)	35 (21.6%)	2 (1.2%)	1 (0.6%)	162
Report[2]	*Exact*	576 (74.6%)	148 (19.2%)	25 (3.2%)	23 (3.0%)	772
	Inexact	653 (95.3%)	6 (0.9%)	5 (0.7%)	21 (3.1%)	685
	Implausible	150 (88.8%)	16 (9.5%)	1 (0.6%)	2 (1.2%)	169

[1] Mean (standard deviation) of the percentage of consistency errors per paper
[2] Frequencies (row percentages) of consistency errors

Table 6
Possible causes of gross consistency errors

Observed[1]	**Expected[2]**	**Report**	**Copy**	**One-two-tailed**	**Precision**	**Unidentified**
Maintain	Reject	Complete	3	7	–	5
		Incomplete	3	5	–	1
Reject	Maintain	Complete	1	–	5	7
		Incomplete	–	–	1	8

[1] Decision that was made
[2] Decision that should have been made

Discussion

When reporting the results of a hypothesis test, APA (2010) recommends including the value of the statistic, its degrees of freedom, the corresponding *p*-value, and a measure of the effect size (see also Wilkinson & TFSI, 1999). Our study assessed if the way of reporting in four Spanish psychology journals indexed in the JCR adjusts to APA regulations. Of the 1,717 tests in our review, 24.1% offer incomplete information. In an additional 5.3% of the tests, in addition to being incomplete, it is impossible to deduce the necessary information from the description of the design. Incomplete information is more frequent for *t* and χ^2 statistics than for *F* statistics (approximately half of the *t* and χ^2 statistics are offered without their corresponding degrees of freedom).

If one considers that a report that does not include a measure of effect size is not a complete report, then only 24.3% of the analyzed tests are complete (i.e., include test statistic, degrees of freedom, *p*-value, and effect size). Only 41% of the reviewed articles include some effect size measure. This value lies close to the lower limit of the interval (40%-73%) that was reported by McMillan and Foley (2011) in educational psychology journals. Reporting the effect size is more frequent with the *t* statistic than with the χ^2 statistic.

In sum, the reviewed articles do not strictly follow APA regulations. We reviewed the instructions given to the authors by the analyzed journals. All of them refer to APA regulations (2001; 2010). For example, the *An Psicol* website includes a summary of these regulations with examples of the information that must accompany each test. The fact that *An Psicol* is the journal for which we observed the highest percentage of incomplete statistics per article (36.4% average) seems to indicate that the problem is not related to dissemination but instead to a lack of knowledge (on the authors' part, of course, but also on the copyeditors' part).

Nevertheless, the main goal of this work was not to review the adherence of the offered information to APA regulations, but to assess the consistency of the statistical results. We observed: (a) that 15.2% of the registered tests contain a consistency error (12.2% of the complete reports); (b) that in 2.8% of the tests a *gross* error is committed (2.3% of the complete reports); and (c) that 62.7% of the reviewed articles includes at least one consistency error (23.5% of the articles include at least one *gross* error; 17.6% if only complete reports are taken into account). Although these percentages may seem high, they are similar to the ones found in other studies (see Bakker & Wicherts, 2011; Berle & Starcevic, 2007; García-Berthou & Alcaraz, 2004). The study by Bakker and Wicherts (2011), for example, shows that the percentage of consistency errors in low impact journals (impact factor lower than 1.5) oscillates between 10.3% and 21.3%.

Consistency errors are more frequent for tests reported with an exact *p*-value (25.4%) than for tests with an inexact *p*-value (4.7%). Bakker and Wicherts' (2011) results are similar: 27.1% and 8.1% in low impact journals. This result is logical: on the basis of inexact *p*-values it is not possible to detect *slight* errors and also some *moderate* errors. However, whether the *p*-value is reported exactly o inexactly does not seem to affect the percentage of *gross* errors.

Consistency errors are also more frequent among incomplete statistics (23.9%) than among the complete ones (12.2%). This might be due to the possibility that the authors who offer incomplete reports are less expert or less careful when reviewing

and/or copying the results from computer programs. However, if we take into account that we completed the incomplete reports with information from the same article, we cannot overlook the fact that, in some cases, the observed inconsistency may have been due to a mistake in the information on which we based our estimation of the degrees of freedom rather than to a mistake in the reported *p*-value.

Concerning the conclusions derived from the reported *p*-values, the majority of errors (81.4%) can be considered not too relevant. However, 18.6% of the errors imply a change in the test conclusion (a change in the rejection or acceptance of the null hypothesis). Furthermore, 23.5% of the reviewed articles would have to change at least one conclusion (17.6% if we focus only on complete reports). This percentage is noticeably larger than the 3% reported by Berle and Starcevic (2007) for psychiatry journals, somewhat higher than the 12% reported by García-Berthou and Alcaraz (2004) for medical journals, and similar to the 19.2% reported by Bakker and Wicherts (2011) for low impact journals.

Bakker and Wicherts (2011) point out that most of the conclusion changes are produced in the direction expected by the researchers (rejecting the null hypothesis when it should in fact be maintained). In our revision we did not observe this bias: of the 46 *gross* errors, in 22 cases the null hypothesis is rejected and in 24 it is maintained (this pattern does not change when focusing only on complete reports).

Although in 45.6% of the *gross* errors we have not been able to identify a possible cause, 26.1% of these errors seem to be due to the use of a two-sided instead of one-sided *p*-value (this refers only to the *t* statistic). That is, although the researchers formulate a directional hypothesis (e.g., it is expected that this will be better than that), they base their decisions on two-sided *p*-values. It is probable that this is due to the fact that computer programs usually offer, by default, the two-sided *p*-value. This type of errors as well as *copy* and *lack of precision* errors could be avoided if researchers more carefully focused their attention when selecting and copying results from computer applications.

Finally, it is important to note that consistency errors can affect the conclusions of meta-analyses. Because meta-analyses focus on magnitudes of effects rather than on statistical significance, the results of meta-analyses can be contaminated to the extent to which the consistency errors bias the estimation of the effect size. According to our results, 9.5% of the gross consistency errors produce deviations larger than 0.1 in the effect size estimations. As Bakker and Wicherts (2011) indicate, these deviations can be considered large enough to importantly affect the results of meta-analyses.

To correctly evaluate the observed consistency errors, one must take into consideration that our review does not include studies in which none of the selected statistics were used (F, t, χ^2). This means that our review does not include studies that use other statistics (that in many cases are more difficult to report and interpret) or studies in which there is a mention of statistical significance without reporting the type of the applied statistics. We must add to this the not uncommon practice of omitting statistical information concerning non-significant results. These types of errors perhaps deserve another study.

In conclusion, concerning the *quality of the information*, 5.3% of the registered statistical results do not include the minimally required information to calculate the *p*-value, and 24.1% do it incompletely (i.e., not explicitly). Also, only 38.2% of the statistical results are accompanied by a measure of effect size.

Concerning consistency errors in complete reports: (a) *slight* errors appear in 8.8% of the reviewed tests, *moderate* errors in 1.1%, and *gross* errors in 2.3%; (b) 17.6% of the reviewed articles contain at least one gross consistency error; (c) 9.5% of the detected errors generate effect size estimations that hold discrepancies that exceed 0.1 points, and the discrepancies derived from consistency errors come to exceed two points.

These results imply a need to improve the way in which statistical results are reported in Spanish psychology journals. Bakker and Wicherts (2011) have made several recommendations that may help to achieve such improvements. We believe that reporting more completely and precisely is not exclusively the task of the researchers. Certainly, researchers must improve their training in methodological issues or turn to consultants who can offer help, but it is also the task of journal editors to choose copy editors who focus attention on methodological issues in a solvent manner.

Acknowledgments

We thank the editorial board of *Psicothema* and two anonymous reviewers for their valuable advice and comments. We also thank Amanda Davies and David Jacobs for his corrections on the drafts of this paper.

References

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

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

American Psychological Association (2010). *Publication manual of the American Psychological Association* (6ª 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.

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.

Botella, J., & Gambara, H. (2002). *Qué es el meta-análisis*. Madrid: Biblioteca Nueva.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2ª ed). New York: Academic Press.

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

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., Lo, J., McMenamin, N., & Wilson, S. (2007). Statistical reform in psychology. Is anything changing? *Psychological Science*, *18*, 230-232.

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

Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

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

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

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.

McMillan, J.H., & Foley, J. (2011). Reporting and discussing effect size: Still the road less traveled? *Practical Assessment, Research and Evaluation*, *16*(14).

Murphy, K.R. (1997). Editorial. *Journal of Applied Psychology*, *82*, 3-5.

Neyman, J., & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, *20A*, 175-240 (1ª parte), 263-294 (2ª parte).

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

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

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

Sánchez-Meca, J., & Botella, J. (2010). Revisiones sistemáticas y meta-análisis. Herramientas para la práctica profesional. *Papeles del Psicólogo*, *31*, 7-17.

Sun, S., Pan, W., & Wang, L.L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, *102*, 989-1004.

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

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

Wagenmakers, E.J. (2007). A practical solution to the pervasive problems of *p*-values. *Psychonomic Bulletin and Review*, *14*, 779-804.

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

Zientek, L.R., Capraro, M.M., & Capraro, R.M. (2008). Reporting practices in quantitative teacher education research: One look at the evidence cited in the AERA panel report. *Educational Researcher*, *37*, 208-216.