Behav Res
DOI 10.3758/s13428-011-0089-5

The (mis)reporting of statistical results in psychology journals

Marjan Bakker · Jelte M. Wicherts

Abstract In order to study the prevalence, nature (direction), and causes of reporting errors in psychology, we checked the consistency of reported test statistics, degrees of freedom, and *p* values in a random sample of high- and low-impact psychology journals. In a second study, we established the generality of reporting errors in a random sample of recent psychological articles. Our results, on the basis of 281 articles, indicate that around 18% of statistical results in the psychological literature are incorrectly reported. Inconsistencies were more common in low-impact journals than in high-impact journals. Moreover, around 15% of the articles contained at least one statistical conclusion that proved, upon recalculation, to be incorrect; that is, recalculation rendered the previously significant result insignificant, or vice versa. These errors were often in line with researchers' expectations. We classified the most common errors and contacted authors to shed light on the origins of the errors.

Keywords Statistical result · Error · *p* value · Significance testing · Expectation

The majority of empirical articles in psychology report numerous statistical results that are the outcome of null hypothesis significance testing, or NHST (Cohen, 1962, 1994; Hubbard & Ryan, 2000; Maxwell, 2004; Sterling et al. 1995). Because these results provide the basis of substantive conclusions and provide the input for meta-analyses, it is important that statistical results should be reported accurately. Nevertheless, there are reasons to expect that some of the statistical results, as presented in articles in psychological journals, will be in error, as evidenced by inconsistencies between the reported *p* value and the test statistic with the accompanying degrees of freedom (*df*).

As for all expert behavior, the reporting of statistical results is subject to human error (Reason, 1990). For instance, the misreporting of a statistical result may be the result of a typo or of misreading the output of a statistical software program. Moreover, misreporting may be caused by the application of incorrect rules or by a lack of knowledge of the statistical test. For example, the total *df* in an ANOVA may be taken to be the error *df* in the reporting of an *F* test, or the researcher may divide the reported *p* value of a χ^2 or *F* test by two, in order to obtain a one-sided *p* value, whereas the *p* value of a χ^2 or *F* test is already a one-sided test.

There are reasons to expect that errors in the reporting of statistical results will be biased toward the researcher's expectations.[1] Several studies have shown that scientists are subject to confirmation bias in analyzing their data; that is, their reaction to empirical results depends on whether these results support their hypotheses (Edwards & Smith, 1996; Fugelsang et al. 2004; Koehler, 1993; Mynatt et al. 1977). For instance, Fugelsang et al. interviewed molecular biologists concerning their reactions to data that were either consistent or inconsistent with their hypothesis. The dominant reaction was to dismiss the inconsistent data on methodological grounds, "while data consistent with a

M. Bakker (✉) · J. M. Wicherts
Department of Psychology, Psychological Methods,
University of Amsterdam,
Roetersstraat 15,
1018 WB, Amsterdam, The Netherlands
e-mail: M.Bakker1@uva.nl

[1] Obviously, errors in the published article may also arise because of errors by others than the researchers—for instance, in the process of typesetting. However, most publishers ask the authors of articles to correct proofs before publication, and so the final responsibility of correct reporting almost always lies with the authors.

Published online: 15 April 2011

theory [were] met with little scrutiny" (Fugelsang et al., 2004, p. 92). If researchers have a preference for a particular result, it is likely that errors that are consistent with this preference are more likely to go undetected than errors that are inconsistent. So, given that the upshot of many psychological studies is determined largely by the outcome of NHST (Cumming et al., 2007; Mahoney, 1977; Rosnow & Rosenthal, 1989) and researchers usually have a preferred outcome—that is, a significant result—we expect errors to favor the preferred outcome.

Barring some earlier work on recording errors in psychology (Rosenthal, 1978; Rossi, 1987), we know of no studies addressing the congruence of statistical results reported in psychology journals. Two studies in related fields did reveal a rather high error rate in the reporting of statistical results. The errors studied concerned the congruence of the test statistic, *df*, and the *p* value. Garcia-Berthou and Alcaraz (2004) checked the congruence in 44 articles published in *Nature* and *British Medical Journal* (*BMJ*) by comparing the reported test statistics and *df* with the reported *p* value. They found that 11.6% of the statistical results reported in *Nature* and 11.1% of the statistical results reported in *BMJ* were incongruent. At least one such error appeared in 38% and 25% of the articles of *Nature* and *BMJ,* respectively. Berle and Starcevic (2007) obtained approximately the same percentages in their study of two psychiatry journals. Specifically, of the statistical results reported in 96 articles in the *Australian and New Zealand Journal of Psychiatry* and *Acta Psychiatrica Scandinavica*, 14.3% were incongruent. In these journals, 36% of the articles with statistical results included at least one error. In both these works and in ours, the focus is on NHST. This method has been extensively criticized (Cohen, 1994; Nickerson, 2000; Wagenmakers, 2007; Wilkinson and Task Force on Statistical Inference, 1999). Notwithstanding these criticisms, NHST remains the most commonly used method of statistical testing in psychology (Cumming et al., 2007). Additional information, such as effect sizes or confidence intervals (CIs), that should supplement NHST are still rarely reported (Cumming et al., 2007; Hoekstra et al. 2006; Vacha-Haase et al. 2000).

The goals of the present article are (1) to establish the prevalence and magnitude of congruence errors for statistical results in psychology articles by recomputing the *p* values as reported in these articles; (2) to establish the prevalence of incompletely reported statistical results (e.g., *F* tests that are reported without the two *df* that characterize the distribution); (3) to document the most common causes of these incongruencies; and (4) to verify whether congruence errors related to NHST are more likely to favor the preferred (alternative) hypothesis.

Psychology journals differ in quality and prestige, as reflected by impact factors (IFs) and rejection rates of submitted manuscripts (Buffardi & Nichols, 1981; Rotton et al. 1993). In the first study, we focused on the number and magnitude of congruence errors (goal 1) and the number of incompletely reported statistical results (goal 2) in all the articles published in 2008 in three randomly selected high-impact and in three randomly selected low-impact psychology journals. Given the differences between these two types of journals in rejection rates and possible quality standards, we expected the articles in the high-impact journals to contain fewer errors than the articles in the low-impact journals. Furthermore, we examined in detail the type (goal 3) and direction (goal 4) of the errors. Our second study served to establish whether the obtained percentages of congruence errors and incompletely reported statistical results generalized to other psychology articles (goals 1 and 2). To this end, we studied congruence errors and incompletely reported statistical results in a random sample of psychology articles published in 2008. In this second study, we contacted all the authors of articles that contained congruence errors, in an attempt to determine the origins of the errors (goal 3). In the last section of our article, we discuss implications and formulate recommendations for improving the practice of reporting *p* values in psychology.

Study 1

Method

To compare high-impact psychology journals with low-impact psychology journals, we used a stratified sampling design. We first obtained the IFs of 447 psychology journals from the JCR social sciences edition of 2007. We then randomly selected three high-impact journals (IF > 4) and three low-impact journals (IF < 1.5) and included in our analysis all the empirical articles published in these journals in 2008. In accordance with the methods employed by Berle and Starcevic (2007), we checked only χ^2, *t*, and *F* tests, because in most null-hypothesis testing, these tests are applied. We did not consider χ^2, *t*, and *F* tests that were used in regression analysis or model fitting, because statistical tests of regressions are often not presented fully and because model fitting (e.g., in structural equation modeling) normally is aimed at *not* rejecting the null hypothesis.

We included both *exactly reported* p *values* (e.g., $p = .034$) and *inexactly reported* p *values* (e.g., $p < .05$). Both are error prone, but in different ways. Because of the wider range of inexact *p* values, fewer errors were expected, but the magnitude of the errors was likely to be greater. Therefore, we took these differences between the exactly and inexactly reported results into account. Note that earlier

studies (Berle & Starcevic, 2007; Garcia-Berthou & Alcaraz, 2004) involved only exactly reported *p* values.

We gleaned from each article the test statistics, *df*, and *p* value. We also recorded whether a one- or two-tailed test was used. Unless stated otherwise, *t* tests were considered two-tailed. We ensured that inconsistencies were not due to Bonferroni correction or similar procedures. We recalculated the *p* value on the basis of the reported test statistic and *df*. Because it is infeasible to recalculate the *p* value associated with incomplete results, these results were considered as missing values and were not taken into account when establishing the congruence error prevalence. These incomplete results were, however, included in our examination of error types. We considered a reported *p* value to be incorrect if it differed from our recalculated *p* value. We recalculated *p* values in *R version 2.9.0*, *Microsoft Office Excel 2003*, and *SPSS 15.0 for Windows* to make sure that our results were consistent over different software packages. The use of different packages showed differences only in the seventh decimal or smaller and so had no bearing on our results. Furthermore, we verified that congruence errors were not due to correct rounding by the original authors. For example, consider a statistical result that is reported as "$t(15) = 2.3$, $p = .033$." Recalculation based on the given *t* value and *df* would give a *p* value of .0362. Nevertheless, in this case, the reported *p* value is considered to be correct because the "true" test statistic could range from 2.25 to 2.35 and, accordingly, the correct *p* value could range from .033 to .040. Therefore, this example would not represent a congruence error.

Because these incongruent statistical results can be used in a meta-analysis, we also wanted to learn about the magnitude and potential influence of the errors. Effect sizes in meta-analyses in psychology often concern the comparison of two groups (e.g., a clinical and a control group) or the relation between two variables (e.g., brain volume and IQ) (Borenstein et al. 2009). Because we did not include relational data in this study, we will focus only on the comparison of two groups and, therefore, will include only errors from *t* tests or *F* tests with one *df* in the numerator. We calculated Cohen's *d* on the basis of the reported *t* value or the square root of the *F* value under the assumption of equal group sizes. This value was subsequently compared with Cohen's *d* as based on a newly calculated *t* value based on the reported *df* and reported *p* value. The absolute mean difference was calculated to get an indication of the potential bias in meta-analytic outcomes due to the incongruence.

We searched all the articles for statistical results and imported them to a separate Excel file by hand. Subsequently, we recalculated the *p* values on the basis of the reported test statistic and *df* and compared these values with the reported *p* values. To counterbalance potential selection, copying, and calculation errors during this process, we carried out the following checks in our analysis. To prevent copying errors, additional information was retrieved from the selected articles, such as the number of decimals reported. In an automated procedure, this additional information was compared with the imported statistical results. If this information did not match, the results in the original article were checked again. Furthermore, to prevent the results from being incorrectly classified as incongruent, all statistical results that were incongruent according to our analyses were examined a second time to avoid copying, selection, and calculation errors on our part. Furthermore, an independent rater, who was blind to the aims of the study, identified and copied 256 statistical results from ten articles randomly chosen from our sample. This rater's results were compared with the results obtained by the first author. The selection of statistical results was consistent in 95.4% of the cases, and the copying of the statistical results was consistent in 99.6% of the cases. The selection discrepancies consisted of three wrongly included regression *F* test and nine values reported in a table. All were correctly reported in the original articles. The only copying error consisted of a *t* value of 0.20, incorrectly copied as 0.21. However, both values are congruent with the reported *p* value. The lack of perfect agreement had no effect of substance on our main results. Since our interest lay with establishing the prevalence of congruence errors, and because all these errors were checked multiple times, we are confident that our own coding errors have no bearing on our main findings. If anything, any additional error on our part will have led to an underestimation of the overall error rate.

Results

Of the 25 high-impact journals, we selected *Journal of Child Psychology and Psychiatry* (JCPP; IF = 4.432), *Development and Psychopathology* (DP; IF = 4.374), and *Journal of Personality and Social Psychology* (JPSP; IF = 4.505). Because JPSP had two 2008 volumes and a rather large number of articles, we restricted our attention to Volume 94. The three randomly selected low-impact journals were *Journal of Black Psychology* (JBP; IF = 0.860), *Journal of Applied Developmental Psychology* (JADP; IF = 1.055), and *Journal of Research in Reading* (JRR, IF = 1.340).[2] All empirical articles published in 2008 were included in our

[2] During the random selection process of the high-impact journals, we also selected *American Psychologist* (IF = 6.987) and *Behavioral Brain Sciences* (IF = 17.462). However, both journals rarely include experimental results and were, therefore, excluded from further analysis. We also selected *Women & Therapy* (IF = 0.080) as a low-impact journal. However, this journal included almost no experimental results and was, therefore, excluded from further analysis.

analyses. We found 4,248 statistical results (2,624 [62%] *F*, 982 [23%] *t*, and 642 [15%] χ^2 tests) in the selected articles, 4,077 (96%) of which were reported completely. Table 1 contains the number of articles, the number of articles that included complete statistical results, the total number of complete statistical results, mean number of statistical results per article, and number of gross errors and errors in each of the selected journals.

The numbers of exactly reported and inexactly reported statistical results are given in Table 2, along with the number of errors and gross errors (as a subset of the errors). It proved infeasible to determine whether authors used a significance level other than .05, since the nominal significance level was often not explicated by the authors. Therefore, an error was recorded as a *gross error* only if the error affected the statistical decision on the basis of the nominal significance level of .05. We found a congruence error in 17.1% of the exactly reported statistical results and in 6.7% of the inexactly reported statistical results. Furthermore, we found a gross error in 1.5% of the exactly reported statistical results and in 1.1% of the inexactly reported statistical results. Table 3 contains the numbers of articles with at least one error or gross error. We found at least one error in 53.7% of the articles with exactly reported *p* values and at least one error in 37.1% of the articles with inexactly reported *p* values. Furthermore, (at least one) gross error was found in 12.4% of the articles with exactly reported *p* values and in 12.4% of the articles with inexactly reported *p* values. Figure 1 provides a flow chart of the categorization of the examined articles.

To understand the severity of the reporting errors, we computed their potential effects on Cohen's *d* metric. We found 147 incongruent results that could be included in a meta-analysis in Cohen's *d* metric (*t* test or *F* test with one *df* in the numerator). The absolute mean difference in Cohen's *d* is 0.174 (median = 0.038, *SD* = 0.564). The difference ranged from 0.0003 to 5.043. Twenty-five percent of these errors were small (i.e., less than .01), but 23% of the differences were greater than .10. This difference was large enough to have a profound effect on the outcome of meta-analyses.

In Tables 2 and 3, the number of errors and gross errors are also reported separately for the high- and low-impact journals. Errors and gross errors are dichotomously scored, and the statistical results are not statistically independent because of the multilevel structure of the data. Therefore, we used multilevel logistic regression utilizing R's lme4 package (Bates & Sarkar, 2007) to compare the prevalence of errors between high- and low-impact journals. In the

Table 2 Number of statistics, errors, and gross errors in high- and low-impact journals

		No. Statistics	No.Errors	No. Gross Errors
High	Exact	961	144 (15.0%)	11 (1.1%)
	Inexact	2,327	147 (6.3%)	26 (1.1%)
	Total	3,288	291 (8.9%)	37 (1.1%)
Low	Exact	207	56 (27.1%)	7 (3.4%)
	Inexact	581	47 (8.1%)	6 (1.0%)
	Total	789	103 (13.1%)	13 (1.6%)
Total	Exact	1,168	200 (17.1%)	18 (1.5%)
	Inexact	2,908	194 (6.7%)	32 (1.1%)
	Total	4,077	394 (9.7%)	50 (1.2%)

Note. The high-impact journals are *Journal of Child Psychology and Psychiatry*, *Development and Psychopathology*, and *Journal of Personality and Social Psychology*. The low-impact journals are *Journal of Black Psychology*, *Journal of Applied Developmental Psychology*, and *Journal of Research in Reading*.

Table 1 Number and percentage of articles with complete statistical results and total number, mean, and standard deviation of complete statistical results

	High			Low		
	JCPP	DP	JPSP	JBP	JADP	JRR
No. articles	119	61	68	22	38	25
No. articles with χ^2, t, or F tests	55 (46.2%)	29 (47.5%)	58 (85.3%)	13 (59.1%)	24 (63.2%)	15 (60.0%)
No. statistical results	798	608	1,882	89	323	377
M	14.51	20.97	32.45	6.85	13.46	25.13
SD	18.35	20.52	24.48	4.32	20.59	24.36
No of errors	71 (8.9%)	30 (4.9%)	190 (10.1%)	19 (21.3%)	45 (13.9%)	39 (10.3%)
No. of gross errors	8 (1.0%)	4 (0.7%)	25 (1.3%)	2 (2.2%)	6 (1.9%)	5 (1.3%)
No. articles with errors	24 (43.6%)	15 (51.7%)	37 (63.8%)	7 (53.5%)	13 (54.2%)	10 (66.7%)
No. articles with gross errors	5 (9.1%)	4 (13.8%)	16 (27.6%)	2 (15.4%)	5 (20.8%)	3 (20.0%)

Note. Gross errors are a subset of errors. *JCCP*, *Journal of Child Psychology and Psychiatry*; *DP*, *Development and Psychopathology*; *JPSP*, *Journal of Personality and Social Psychology*; *JBP*, *Journal of Black Psychology*; *JADP*, *Journal of Applied Developmental Psychology*; *JRR*, *Journal of Research in Reading*

Table 3 Number of articles with statistics, number of articles with at least one error, and number of articles with at least one gross error in high- and low-impact journals

		No.Articles With Statistics	No.Articles With Errors	No. Articles With Gross Errors
High	Exact	92	47 (51.1%)	10 (10.9%)
	Inexact	131	46 (35.1%)	17 (13.0%)
	Total	142	76 (53.5%)	25 (17.6%)
Low	Exact	29	18 (62.1%)	5 (17.2%)
	Inexact	47	20 (42.6%)	5 (10.6%)
	Total	52	30 (57.7%)	10 (19.2%)
Total	Exact	121	65 (53.7%)	15 (12.4%)
	Inexact	178	66 (37.1%)	22 (12.4%)
	Total	194	106 (54.6%)	35 (18.0%)

Note. See note to Table 2 for the titles of high- and low-impact journals.

model, each test statistic is nested within an article, while each article is nested within a journal. We started with an empty model with only article and journal as random variables to predict the errors (AIC = 2,329.2; BIC = 2,348.1; LogLik = –1,161.6). Next, we added a fixed effect of exact versus inexact at the statistic level to indicate the way the *p* value was reported. This improved model fit substantially (AIC = 2,291.5; BIC = 2,316.8; LogLik = –1,141.8), $\chi^2(1) = 39.68$, $p < .001$, which is to be expected because a range of *p* values in inexactly reported statistical results is less likely to be incongruent than an exactly reported *p* value. Next, we added impact of the journal (high vs. low) at the journal level as a fixed effect. This also improved model fit considerably (AIC =2,286.5; BIC 1= 2,318.1; LogLik = –1,138.3), $\chi^2(1) = 6.99$, $p = .008$. An interaction effect between impact and exact/inexact did not improve the model fit (AIC = 2,287.9; BIC = 2,325.8; LogLik = –1,138.0), $\chi^2(1) = 0.58$, $p = .447$. The final model gave the following results. The variance of article was 1.46, and the variance of journal was 0.00 in the final model; the intercept was equal to –2.53 (95% CI [–3.00, –2.06]). We found a significant (fixed) effect of exact versus inexact reporting ($\hat{\beta} = 0.93$, 95% CI [0.64, 1.21], $Z = 6.38$, $p < .001$) and of journal's impact ($\hat{\beta} = –0.74$, 95% CI [–1.27, –0.20], $Z = 2.70$, $p = .007$). Thus, more errors were found with exactly reported statistical results, and more errors were found in low-impact journals.

We also modeled the gross errors, starting with an empty model with only article and journal as random variables (AIC = 539.1; BIC = 558.0; LogLik = –266.5). We added a fixed effect of exact/inexact at the statistics level, but this did not improve the model fit (AIC = 539.4; BIC = 564.7; LogLik = –265.7), $\chi^2(1) = 1.67$, $p = .197$. Also, adding a fixed effect of journal's impact (AIC = 539.4; BIC = 564.7; LogLik = –265.7), $\chi^2(1) = 1.66$, $p = .198$, did not improve the model fit, as compared with the first model. Therefore, we found no statistically significant difference in the proportion of gross errors between exactly and inexactly

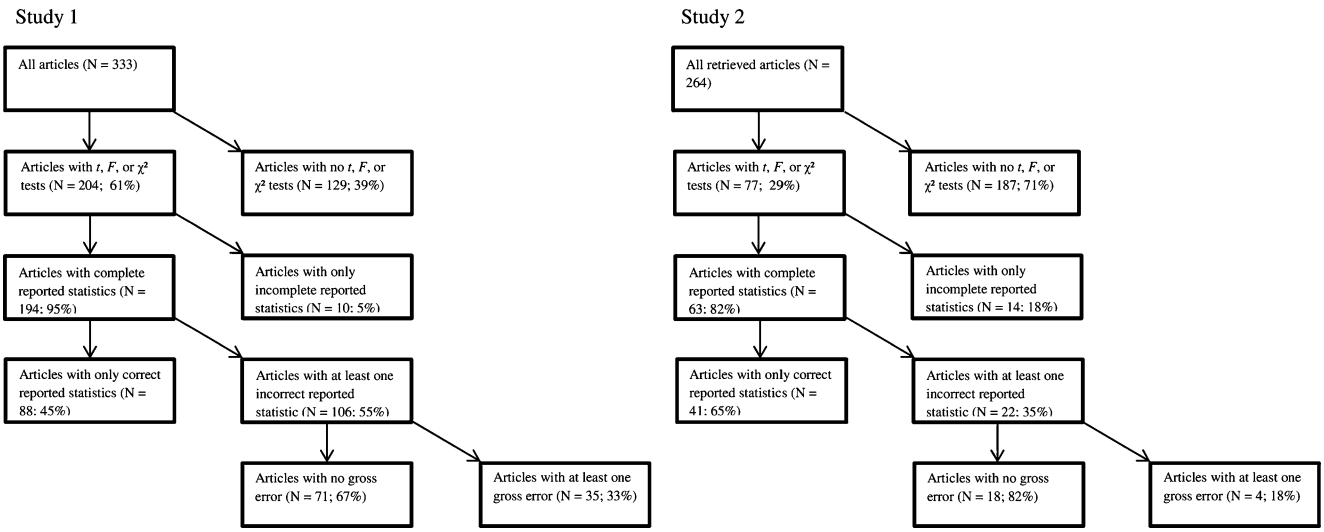

Fig. 1 Flow chart of articles in Studies 1 and 2

reported statistical results and between high- and low-impact journals.

We conducted a further inspection of the errors to get a better understanding of the different error types. Because exactly and inexactly reported statistical results differ in terms of how they can be misreported and also how we were able to detect reporting errors, we classified errors made with respect to exactly and inexactly reported statistical results separately. The errors with respect to exactly reported p values were classified as follows:

1. Incomplete: test statistic, *df*, or p value missing.
2. Rounding errors: wrongly rounded upward or downward. An example of the latter is $F(3, 58) = 2.78$, $p = .04$, while the recalculated p value equalled .04938.
3. Usage of one-sided t tests without a mention of the one-sidedness of the test.
4. Incorrect reporting of the smallest p values. For example, $F(2, 20) = 15.2$, $p = .001$ was reported, which is incongruent, since the correct p value is .000097, which should have been reported as $< .001$ according to the guidelines of the APA Publication Manual (American Psychological Association, 2010).
5. Wrong use of tests, such as dividing the p value of an F or χ^2 test by two to report a one-sided p value, whereas the F or χ^2 test is already a one-sided test. Note that this procedure could be correct for particular F tests that can be transformed to a t test or for a χ^2 test that can be transformed to a Z value because of equivalency.
6. Unidentifiable: the error could not be classified on the basis of the reported information.

The errors with respect to inexactly reported p values were classified as follows:

1. Incomplete: test statistic, *df*, or p value missing.
2. Reported "< .000."
3. Reported "<", when "=" would be correct. For example, $\chi^2(4) = 12.63$, $p < .01$ is reported, whereas the correct p value is .0132, which could be reported as $p = .01$.
4. Wrong use of tests as described under (5) above.
5. Unidentifiable: the error could not be classified on the basis of the reported information.

The numbers of errors per category in the high- and low-impact journals are presented in Table 4 and in Fig. 2. We found that, in total, 171 (4%) statistical results were reported incompletely. Furthermore, the source of many errors was unidentifiable. The impossible tests category included a one-tailed χ^2 test and 2 one-tailed F tests. A sizable portion of the errors with inexactly reported p values were of the "< instead of =" type. Furthermore, many errors among the exactly reported p values appeared to be attributable to incorrect rounding.

Besides these error categories, we investigated the occurrence of what we call *copy–paste errors*. Sometimes a reported test statistic with accompanying *df* and p value appeared to serve as a template in reporting other statistical results (later in the same text). Subsequently, the researcher may have forgotten to edit (some part of) the copied–pasted results, which resulted in a congruence error. The following substantively adapted but otherwise real quotations[3] illustrates this kind of error: "The main effect of ability was significant in the first and second setting, $F(1, 39) = 6.646$, $p = .015$, and $F(1, 26) = 1.175$, $p = .020$." These results were copied to the next paragraph: "As predicted, the main effect of training was significant in the first and second setting, $F(1,39) = 6.646$, $p = .015$ and $F(1,26) = 4.175$, $p = .020$." These results are almost exactly the same, except for the test statistic of the second setting (1.175 vs. 4.175). On top of the copying, all the results are incongruent. Nevertheless, copy–paste errors are not necessarily incongruent. That is, a congruent result may be copied and not altered at all. As a result, this copy–paste error may still be congruent and, therefore, not discovered in our analyses. To obtain a rough indication of the prevalence of copy–paste errors, we searched all the articles for a repetition of the same test statistic and *df*. Exact matches were found in *JCPP* 12 times (1.5% of the statistical results), in *DP* 9 times (1.5%), in *JPSP* 23 times (1.2%), in *JADP* 8 times (2.5%), and in *JRR* 4 times (1.1%). This suggests that copy–paste errors are quite common. In addition, these particular errors suggest the existence of even more errors than can be inferred from our analyses on the basis of incongruencies.

To investigate whether congruence errors were in the direction of the researchers' hypotheses, we first categorized the exactly reported statistical results as significant results ($p \leq .05$) and nonsignificant results ($p > .05$) and then inspected the errors and gross errors with a multilevel logistic regression model. The total number of statistical results and the number of errors per significance category are reported in Table 5. We started with an empty model with only a random effect of article and journal to predict the errors (AIC = 946.5; BIC = 961.7; LogLik = –470.2). Adding a fixed effect of significant/nonsignificant results at the statistics level did not improve model fit (AIC = 948.5; BIC = 968.7; LogLik = –470.2), $\chi^2(1) = 0.02$, $p = .886$, which means that we did not find a statistically significant difference in the distribution of errors over the significant and nonsignificant results. For the gross errors, we also started with an empty model with only a random effect of article and journal (AIC = 186.0; BIC = 201.2; LogLik = –90.00). Including a fixed effect of significant/nonsignificant

[3] We did not cite the source of this error because we do not wish to incriminate the authors.

Table 4 Error categories of exactly and inexactly reported statistical results in high- and low-impact journals

Exactly Reported			Inexactly Reported		
	Category	*n* (%)		Category	*n* (%)
High	1: Incomplete	60 (29.4%)	High	1: Incomplete	72 (32.9%)
	2: Rounding error	60 (29.4%)		2: < 0.0	1 (0.5%)
	3: One-sided *t* tests	3 (1.5%)		3: < instead of =	93 (42.5%)
	4: Lowest *p* values	14 (6.9%)		4: Impossible tests	3 (1.4%)
	5: Impossible tests	0 (0.0%)		5: Unidentified	50 (22.8%)
	6: Unidentified	67 (32.8%)			
Low	1: Incomplete	3 (5.1%)	Low	1: Incomplete	36 (43.4%)
	2: Rounding error	6 (10.2%)		2: < 0.0	3 (3.6%)
	3: One-sided *t* tests	3 (5.1%)		3: < instead of =	27 (32.5%)
	4: Lowest values	7 (11.9%)		4: Impossible tests	0 (0.0%)
	5: Impossible tests	0 (0.0%)		5: Unidentified	17 (20.5%)
	6: Unidentified	40 (67.8%)			

Note. See note to Table 2 for the titles of high- and low-impact journals.

results at the statistics level improved model fit (AIC = 180.1; 200.3; LogLik = –86.0), $\chi^2(1) = 7.95$, $p = .005$. The variation due to article in this final model was estimated at 7.30, and the variation due to journal was 0.00. The intercept was equal to –7.34 (95% CI [–9.11, –5.56]). The fixed effect of significant versus nonsignificant ($\hat{\beta} = 1.81$; 95% CI [0.14, 3.49]; Z = 2.12, $p = .034$) shows that more gross errors were made in the "less than .05" category than in the "greater than .05" category. In other words, gross errors more often rendered a nonsignificant result significant than vice versa. The results are included in Fig. 3. The dots above the main diagonal represent the errors in which the reported *p* value underestimated the actual *p* value, whereas the dots below the main diagonal represent the reported *p* values that overestimated the actual *p* value. The dots in the upper left block represent the instances in which a nonsignificant result was presented as being significant. The dots in the lower right corner represent the gross errors in which a nonsignificant result was presented that turned out to be significant in our recalculation. Furthermore, 31 of the 32 inexactly reported statistical results that contain a gross error report a nonsignificant result as significant.

In addition, we examined all the articles with rounding errors. Of all the articles with rounding errors with a reported *p* value of .05, the reported *p* value was lower than the recalculated *p* value (5 out of 5). Thus, all these rounding errors lead to a significant result. In contrast, in only 12 out of 28 articles with rounding errors concerning other *p* values, at least one reported *p* value was lower than the recalculated *p* value ($p = .044$, two-tailed Fisher's exact test, Cramer's $V = 0.410$). Although it proved infeasible to determine for each statistical test whether or not these were in line with the researchers' expectations, the use of significance testing is normally aimed at rejecting the null hypothesis. Therefore, we interpret these findings as an indication of expectancy effects.

Study 2

The goal of the second study was to establish the generality of the prevalence of congruence errors established in Study 1 by checking errors in a representative sample of articles published in psychology.

Method

We randomly selected 300 articles. To this end, we retrieved from the database all peer reviewed articles published in 2008 that are included in the PsycINFO database. Next, we numbered all of these records, drew a random number on the basis of a uniform distribution for all records, ordered all records accordingly, and selected the first 300 records so ordered. This procedure ensures a fully random sample and, hence, representativeness. The selected articles were handled in the same way as in Study 1. In addition, we recorded whether a CI or a standardized effect size measure (e.g., Cohen's *d* or η^2) was reported in the same paragraph as the results. We computed the percentage of congruence errors in all reported statistical results and the percentage of articles with at least one congruence error. In addition, since standard 8.14 of the ethical standards of the American Psychological Association (2010) states that data should be shared after research results are published, we contacted all authors whose articles included a congruence error and asked them whether they were willing and able to send us, within a period of 2 weeks, a data file with which we could replicate one of their incongruent statistical results. We implemented the time limit on the basis of our previous experiences with data-sharing requests in which a promise to share data "as soon as possible" by the original researchers often meant that we did not hear from them again. We assured all the researchers whom we contacted full confidentiality. Given a nonresponse, we sent a reminder after 1 week.

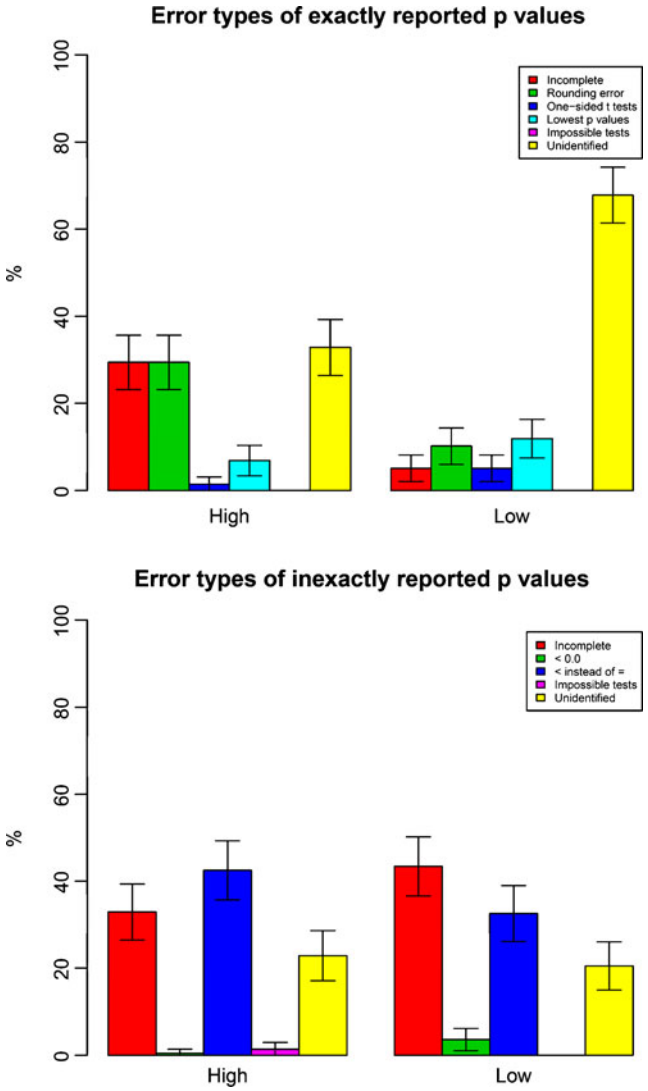

Fig. 2 Overview of the different error categories broken down by high- and low-impact journals and by exactly and inexactly reported statistical results. Wald's confidence intervals are represented in the figure by the error bars attached to each column

Table 5 Number of statistics, errors, and gross errors per significance category in high- and low-impact journals for exactly reported statistical results

		No. Statistics	No. Errors	No. Gross Errors
High	$p \leq .05$	533	89 (16.7%).	10 (1.9%)
	$p > .05$	428	55 (12.9%)	1 (0.2%)
Low	$p \leq .05$	113	33 (29.2%)	5 (4.4%)
	$p > .05$	94	23 (24.5%)	2 (2.1%)
Total	$p \leq .05$	646	122 (18.9%)	15 (2.3%)
	$p > .05$	522	78 (14.9%)	3 (0.6%)

Note. See note to Table 2 for the titles of high- and low-impact journals.

Results

The PsycINFO database contained 88,180 psychology articles that were published in 2008 in a peer-reviewed journal. Because of lack of access to the PDFs within our university's library, we were able to retrieve only 264 articles of our sample of 300. Of these 264 articles, 97 (37%) reported no statistics at all (e.g., because they were book reviews, editorials, or comments). Of the remaining 167 articles, we selected the 77 articles that contained χ^2, *t*, or *F* tests. Many of the unselected articles contained other statistical analyses (e.g., regression analysis, structural equation modelling, nonparametric tests), while 29 articles (17%) reported only *p* values. We did not consider statistical results that were reported only as *p* values, because these could not be verified by our method, since the type of the test that was used was not mentioned. Of the 77 articles with χ^2, *t*, or *F* tests, 63 articles (82%) contained results that were complete. A graphical representation of the selected articles is given on the right-hand side of Fig. 1. We found a total of 809 statistical results (506 [63%] *F*, 223 [28%] *t*, and 80 [10%] χ^2 tests) in the selected articles, 643 (79%) of which were reported completely. The average number of complete statistical results reported per article was 10.21 (*SD* = 9.31). We found congruence errors in 19.4% of the exactly reported statistical results and in 7.5% of the inexactly reported statistical results. These percentages lie close to those in study 1: 17.1% and 6.7%, respectively. Thirty-five percent of the articles contained at least one error. When we combined the incongruent results with the incomplete results, we found at least one misreported result in 55% of the articles.We found a total of seven (1.1%) gross errors in four articles (6.3%). These results, broken down by exactly and inexactly reported statistical results, are presented in Tables 6 and 7.

We found 22 errors with results that could be included in a meta-analysis in Cohen's *d* metric (*t* test or *F* test with one *df* in the numerator). The absolute mean difference in Cohen's *d* due to misreporting was 0.169 (median = 0.065, *SD* = 0.211). The differences ranged from 0.006 to 0.707. Of these differences, 18% can be classified as small (i.e., less than .01), but 41% were substantial (i.e., greater than .10). Furthermore, we checked whether effect size (ES) measures or CIs were reported in the articles of our sample. In only 16 articles (21%) was an ES measure reported, and in only 6 articles (8%) was a CI reported. These results show that NHST continues to be used in ways that defy the guidelines as proposed by the Task Force on Statistical Inference (Wilkinson and Task Force on Statistical Inference, 1999) and the last two versions of the APA Publication Manual (American Psychological Association, 2001, 2010).

We contacted the authors of 21 articles that contained a congruence error. Four of these articles contained a gross

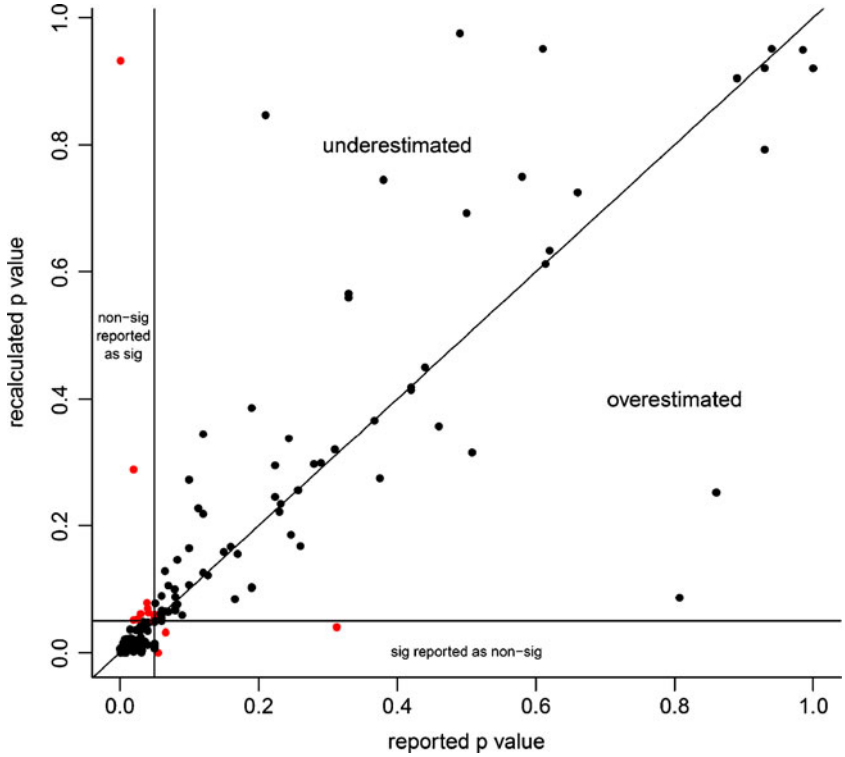

Fig. 3 Erroneously reported *p* values compared with the recalculated *p* value. Correctly reported *p* values would be situated on the diagonal. The dots above the diagonal represent statistical results that were reported as a *p* value lower than the actual *p* value. The dots below the diagonal represent statistical results that were reported as a *p* value higher than the actual *p* value. The dots in the left upper block represent gross errors in which a nonsignificant result is reported as significant. The dots in the right lower block represent gross errors in which a significant result is reported as nonsignificant

error, of which one reported a nonsignificant result that appeared significant after recalculation. The first week we received 11 responses (a 52% response rate). All these quick responses were from authors of articles that did not contain a gross error (p = .035, two-tailed Fisher's exact test, Cramer's V = 0.509). After sending a reminder, the responses totalled 17 responses, 4 of which were from researchers who had committed a gross error. Therefore, the difference in response rate between the researchers with a gross error and those without was no longer significant (p = .546, two-tailed Fisher's exact test, Cramer's V = 0.235). Nevertheless, a nonparametric Mann–Whitney test showed that the authors of articles without a gross error responded considerably faster (median = 3.00 days) than the authors of articles that did contain a gross error (median = 7.00 days; U = 7.50, p = .034). This suggests that the time to respond to a data-sharing request is positively associated with the severity of the statistical error.[4]

Only 5 of the 17 responses included the raw data. However, another six respondents did include a description of the results of a reanalysis of the original data. In addition, six authors responded with some explanation, although these explanations seem to be based only on an inspection of the reported results, rather than on an actual reanalysis. These "unfounded" explanations converged with our own classifications: typos, rounding errors, and the inaccurate use of "<" and "=". The reanalyses by the original authors and our own reanalyses of the raw data revealed that several errors were caused by incorrect reporting of the test statistic or *df*. For example, one author reported a corrected F test but reported the uncorrected *error df* , and another author reported a t test with a *df* of 1, instead of a correct *df* of 38. Another gross error was attributed by the authors to the use of a p value based on one version of the dataset and the use of a test statistic and *df* that were based on a former version of the dataset.

General discussion

We studied the accuracy of the reporting of statistical results in a random selection of high- and low-impact psychology journals (Study 1), and in a fully random sample of recent psychology articles, in which the researchers had employed NHST (Study 2). We found that between 17% (Study 1) and 19% (Study 2) of the exactly

Table 6 Number of statistics, errors, and gross errors in the second study

	No. Statistics	No. Errors	No. Gross Errors
Exact	283	55 (19.4%)	0 (0.0%)
Inexact	360	27 (7.5%)	7 (1.9%)
Total	643	82 (12.8%)	7 (1.1%)

[4] The use of a parametric test did not change the results (M = 3.77, SD = 3.03 vs. M = 7.25, SD = .50), $t(15)$ = 2.24, p = .041.

Table 7 Number of articles with statistics, number of articles with at least one error, and number of articles with at least one gross error in the second study

	No. Articles With Statistics	No. Articles With Errors	No. Articles With Gross Errors
Exact	43	12 (27.9%)	0 (0.0%)
Inexact	52	12 (23.1%)	4 (7.7%)
Total	63	22 (34.9%)	4 (6.3%)

reported statistical results and between 7% (Study 1) and 8% (Study 2) of the inexactly reported statistical results reported in psychological articles are incongruent. These results reveal that the problem of incongruent statistical results is greater in psychology journals than in the other fields that have been studied thus far. In the studies of Garcia-Berthou and Alcaraz (2004) and Berle and Starcevic (2007) of the prevalence of congruence errors in *Nature*, the *British Medical Journal*, and two psychiatry journals, between 11% and 14% of published statistical results with exactly reported *p* values were reported incorrectly. Furthermore, we found that 55% of the articles in the first study and 35% of the articles in the second study contained at least one such error. Moreover, around 1% of the examined statistical conclusions were not supported by the reported test statistic and *df*. More important, we came across at least one unsupported statistical conclusion in 39 of the 257 articles (15%) that we scrutinized in our two studies. In other words, despite passing the peer reviewers, in roughly 1 out of 7 articles in psychology, at least one statistical conclusion appears to have been unfounded on the basis of the presented test results alone.

Moreover, 4% of the statistical results in the first study and 21% of the statistical results in the second study were not completely reported, which goes against the guidelines of the APA Publication Manual (American Psychological Association, 2010). The percentage of incompletely reported results in the psychological literature is even larger, because in our representative sample of psychology articles, we came across 29 articles (17%) in which the statistical results were reported only by a *p* value.

In addition, the results of the first study showed that articles published in low-impact journals contained relatively more congruence errors than articles published in high-impact journals. However, we found no difference between high- and low-impact journals in the prevalence of gross errors. Although the number of statistical results in the first study is large, we examined only three high-impact and three low-impact journals. Therefore, the conclusions about differences between high- and low-impact journals can be dependent on the specific journals included in our study. Despite this potential limitation on the generalizability to other journals, we have no reasons to believe that the journals we selected are unrepresentative for psychology journals with high- and low-impact factors, respectively. In fact, the findings of the second study on the basis of a random (and hence representative) sample of psychological articles do attest to the generality of reporting error frequencies.

Because statistical results from articles can be used for meta-analyses, it is important that results are correctly reported or, at least, that the magnitude of these errors is small. We operationalized the magnitude of reporting errors on the basis of results from *p* values that may feature in meta-analyses with Cohen's *d* and found that the average magnitude of these errors to be substantial (average $d = 0.17$). Reporting results with effect sizes would decrease the unhealthy focus on the significance boundary. However, the second study showed that effect sizes are reported only in around 20% of the articles. Despite many efforts to change reporting practices in psychology (see, e.g., Wilkinson and Task Force on Statistical Inference, 1999), the preponderance of published articles still lack effect sizes. So if *p* values are used, the common misreporting of these *p* values could bias meta-analytic results considerably. The practice of only reporting *p* values as we documented in 17% of the empirical articles in Study 2 should therefore be avoided.

In the second study, we found a similar prevalence of congruence errors as in the first study, although in the second study we came across fewer articles with at least one congruence error than in the first study. This may be due to the fact that the articles in the second study contained fewer statistical results, on average. Especially, the high-impact journals in the first study contained many statistical results per article, mostly because of the common practice of including more than one study per article. Furthermore, we found substantially more incompletely reported statistical results in our second study. Twenty-two percent of the statistical results were not reported according to the guidelines of the APA Publication Manual (American Psychological Association, 2010). This difference between Studies 1 and 2 was probably caused by the overrepresentation of high-impact journals in the first study. For instance only 3 out of 1,882 statistical results in *JPSP* were reported incompletely. This suggests that journal policies can make a difference. The second study involved a fully random sample of articles published in 2008 in peer-reviewed psychology journals, and although the sample of articles may not be large, our results are based on a large number of statistical test results and show clear consistency. Therefore, it is safe to conclude that the prevalence of misreporting (both congruence errors and incomplete results) within psychological articles with statistical results is indeed close to 30%, and

that more than half of the these articles contains at least one such error in the reporting of statistical results.

The present work gives some insight into the types of errors made. To begin with, incompletely reported results are quite common. Most often, a test statistic was given without the mention of one or more *df*s. Another common error is the confusion of "<" with "=". In several articles, we found that inequality and equality signs were used as if they were interchangeable. Furthermore, we came across the wrong use of tests (e.g., F and χ^2 tests from which the p values are divided by two without sound argumentation), problems with the reporting of the smallest p value (e.g., reporting a p value as < .000 or reporting $p = .001$ when $p <$ .001 would have been correct), and rounding errors, and we found evidence of a substantial occurrence of copy–paste errors. We propose recommendations to avoid the misreporting of p values below.

Many congruence errors could not be classified on the basis of the reported information, and so the source of these errors remained unclear. We simply do not know whether the test statistic, the *df*, and/or the p value were misreported. In addition, since we focused only on incongruently reported results, we did not consider other errors in the reporting of statistical results that did not result in incongruency. Thus, it is possible that additional errors may be present, which would surface only following a complete reanalysis of the raw data.

To obtain a better understanding of the origins of the errors made in the reporting of statistics, we contacted the authors of the articles with errors in the second study and asked them to send us the raw data. Regrettably, only 24% of the authors shared their data, despite our request being quite specific and our assurances that the authors would remain anonymous. The degree of nonresponse was in line with the previous results of Wicherts et al. (2006). They requested data from 141 authors of articles in APA journals and observed a response rate of 27%. Nevertheless, some authors, who appeared willing to share their data with us, conducted a reanalysis themselves and informed us of their results. Both the raw data and the results of the reanalyses revealed some additional sources of error. Especially, incongruencies caused by reporting the wrong test statistic or *df* were revealed. Furthermore, several contacted authors gave us more background information on the causes of the incongruencies. For example, one author told us that the reported p value was based on one dataset and that the test statistic and *df* were based on a different, former dataset, which contained an incorrect value. This can be seen as a special case of the copy–paste error with a former result that is only partly edited. Nevertheless, even with access to the raw data, the causes of some errors remained unknown. Given access to the raw data, it is at least possible to determine the correct statistical results, which can be used, for instance, in meta-analyses.

Of special interest was the direction of the congruence errors. Researchers often have specific preferences regarding their results, which may affect the extent to which researchers scrutinize errors in line with or contradicting their preferred results. We hypothesized that congruence errors would more often be in favor of the researchers' expectations. The direction of the gross errors in the first study revealed that 46 of the 50 congruence errors resulted in a significant result. Furthermore, the rounding errors with a p value of .05 were all in favor of the researchers' hypotheses—that is, the alternative rather than the null hypotheses. These errors may have been the result of sloppiness, so they should not be taken to mean that researchers were trying to present a more convincing story than the data could support (Friedlander, 1964). Nonetheless, these results point to the importance of studying further the potential influence of researchers' expectations on the outcome and reporting of their data analyses.

Recommendations

To arrive at more accurate reporting of statistical results, we make the following recommendations.

1. To prevent the inaccurate use of inexact and exact p values, authors and editors should follow the newly revised APA Publication Manual (American Psychological Association, 2010) more closely. The newly revised manual is clear on the reporting of p values: "When reporting p values, report exact p values (e.g., $p = .031$) to two or three decimal places. However, report p values less than .001 as $p < .001$" (p. 114). Note that this guideline applies to both significant and nonsignificant statistical results. This guideline may help to avoid rounding errors and has the additional advantage that the reported results can be more easily verified.
2. To be more informative and to prevent an unhealthy focus on the significance boundary, statistical results should be accompanied by effect sizes and CIs when possible, as is also recommended by the APA Publication Manual (American Psychological Association, 2010): "However, complete reporting of all tested hypotheses and estimates of appropriate effect sizes and confidence intervals are the minimum expectations for all APA journals" (p. 33). This edition of the manual even specifies a reporting format for confidence intervals: "$t(177) = 3.51$, $p < .001$, $d = 0.65$, 95% CI [0.35, 0.95]" (p. 117).

3. Results should be checked by both the (co)author(s) and the reviewers for completeness of the reported statistical results. Specifically, test statistics should always be accompanied by the correct *df*.
4. Sound statistical reviewing is needed to prevent the use of impossible statistical tests, such as incorrectly dividing *p* values of *F* or χ^2 tests. The use of one-sided tests should always be mentioned in the text (preferably next to the test statistic).
5. Researchers should be aware that the use of copy–paste in statistical reporting is error prone. Possible copy–paste errors should be checked during the copy-editing process.
6. Raw data should be made available as a matter of principle—not only to check for possible errors in the reporting of statistical results, but also to have complete and correct statistical information to perform later meta-analyses.

To set an example, we have employed these checks in the present work in the hope that its quality and validity will in no way be affected negatively by our own fallibility. In addition, the raw data from both our studies are available upon request.

Author Note The preparation of this article was supported by Grants 451-07-016 and 400-08-214 from the Netherlands Organization for Scientific Research (NWO).

References

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

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

Bates, D., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes (R package Version 0.999375-32) [Computer software]. Retrieved from http://cran.r-project.org/web/packages/lme4/index.html

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

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, U.K.: Wiley.

Buffardi, L. C., & Nichols, J. A. (1981). Citation impact, acceptance rate, and APA journals. *American Psychologist, 36*, 1453–1456. doi:10.1037/0003-066X.36.11.1453

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145–153. doi:10.1037/h0045186

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist, 49*, 997–1003. doi:10.1037/0003-066X.49.12.997

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., et al. (2007). Statistical Reform in Psychology. *Psychological Science, 18*, 230–232. doi:10.1111/j.1467-9280.2007.01881.x

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology, 71*, 5–23. Retrieved from http://psycnet.apa.org/journals/psp/

Friedlander, F. (1964). Type I and Type II bias. *American Psychologist, 19*, 198–199. Retrieved from http://psycnet.apa.org/journals/amp/

Fugelsang, J. A., Stein, C. B., Green, A. E., & Dunbar, K. N. (2004). Theory and data interactions of the scientific mind: Evidence from the molecular and the cognitive laboratory. *Canadian Journal of Experimental Psychology, 58*, 86–95. Retrieved from http://psycnet.apa.org/journals/cep/

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

Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of *p* values. *Psychonomic Bulletin & Review, 13*, 1033–1037. Retrieved from < Go to ISI>://000245529200015.

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology—and its future prospects. *Educational and Psychological Measurement, 60*, 661–681. doi:10.1177/0013164400605001

Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes, 56*, 28–55. doi:10.1006/obhd.1993.1044

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*, 161–175. doi:10.1007/BF01173636

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*, 147–163. doi:10.1037/1082-989X.9.2.147

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: Experimental-study of scientific inference. *Quarterly Journal of Experimental Psychology, 29*, 85–95. doi:10.1080/00335557743000053

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241–301. doi:10.1037//1082-989x.5.2.241

Reason, J. (1990). *Human error*. Cambridge: Cambridge University Press.

Rosenthal, R. (1978). How often are our numbers wrong? *American Psychologist, 33*, 1005–1008. doi:10.1037/0003-066X.33.11.1005

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

Rossi, J. S. (1987). How often are our statistics wrong? A statistics class exercise. *Teaching of Psychology, 14*, 98–101. doi:10.1207/s15328023top1402_8

Rotton, J., Levitt, M. J., & Foos, P. (1993). Citation impact, rejection rates, and journal value. *American Psychologist, 48*, 911–912. doi:10.1037/0003-066X.48.8.911

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician, 49*, 108–112. Retrieved from http://www.jstor.org/stable/2684823

Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., Lance, T. S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory & Psychology, 10*, 413–425. doi:10.1177/0959354300103006

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*, 779–804. Retrieved from < Go to ISI>://000251227600001.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*, 726–728. doi:10.1037/0003-066x.61.7.726

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604. doi:10.1037/0003-066X.54.8.594