The (non-)significance of reporting errors in economics: Evidence from three top journals

Peer-reviewed author version

Putz, P & BRUNS, Stephan (2021) The (non-)significance of reporting errors in economics: Evidence from three top journals. In: Journal of economic surveys (Print) , 35 (1), p. 348-373.

DOI: 10.1111/joes.12397
Handle: http://hdl.handle.net/1942/32738

The (Non-)Significance of Reporting Errors in Economics: Evidence from Three Top Journals

Peter Pütz, Stephan B. Bruns[*]

September 8, 2020

Abstract

We investigate the prevalence and sources of reporting errors in 30,993 hypothesis tests from 370 articles in three top economics journals. We define reporting errors as inconsistencies between reported significance levels by means of eye-catchers and calculated *p*-values based on reported statistical values, such as coefficients and standard errors. While 35.8% of the articles contain at least one reporting error, only 1.3% of the investigated hypothesis tests are afflicted by reporting errors. For strong reporting errors for which either the eye-catcher or the calculated *p*-value signals statistical significance but the respective other one does not, the error rate is 0.5% for the investigated hypothesis tests corresponding to 21.6% of the articles having at least one strong reporting error. Our analysis suggests a bias in favor of errors for which eye-catchers signal statistical significance but calculated *p*-values do not. Survey responses from the respective authors, replications and exploratory regression analyses indicate some solutions to mitigate the prevalence of reporting errors in future research.

Keywords: Reporting errors; reproducibility; replications; questionable research practices

JEL codes: A11, B40, A12

Acknowledgments: We would like to thank Guido Bünstorf, Chris-Gabriel Islam, Elmar Spiegel and participants at

[*] Corresponding author: Stephan Bruns, Chair of Econometrics, Department of Economics, University of Goettingen, Humboldtallee 3, 37073 Goettingen, stephan.bruns@uni-goettingen.de

MAER-Net 2019, DAGStat 2019, Statistical week 2019 and seminars in Bielefeld, Hasselt and Göttingen for helpful feedback. Stephan Bruns is grateful for funding from the German Research Foundation (DFG) under the project „Replications in Empirical Economics: Necessity, Incentives and Impact“ (project number: 405039391).

1 Introduction

The reliability of empirical research is subject to intensive debate (e.g. Munafò et al., 2017; Wasserstein et al., 2016), with economics being no exception (e.g. Vivalt, 2019; Bruns et al., 2019; Brodeur et al., 2016; Doucouliagos et al., 2018; Chang and Li, 2017). Most prominently, Ioannidis et al. (2017) find evidence of an inflation of significant *p*-values suggesting *p*-hacking (Simmons et al., 2011; Bruns and Ioannidis, 2016; Leamer, 1983), HARKing (Kerr, 1998) and publication bias (Franco et al., 2014; Rosenthal, 1979) to be common practices in empirical economics, as has been shown for many other disciplines (e.g. Albarqouni et al., 2017; O'Boyle et al., 2017; Gerber and Malhotra, 2008a,b). However, reported significance levels and statistical values are usually assumed to be correct and little research has addressed the rate of errors in reported findings. In this paper, we investigate the prevalence of reporting errors in three top economics journals and shed light on potential sources.

We define reporting errors as inconsistencies between reported levels of statistical significance by means of eye-catchers (mostly stars) and calculated *p*-values based on reported statistical values, such as coefficients and standard errors. Errors in reporting may result from honest mistakes originating, for instance, from manually transferring empirical findings from statistical software to word processing software, from accidental mistakes in rounding or during typesetting and insufficient proofreading by the authors. Errors may also result from questionable research practices such as intentionally rounding down *p*-values to let them appear statistically significant (e.g., John et al., 2012; Wicherts et al., 2011). It is also common in economics to report many regression models in one table to convince the reader of the robustness of the main findings and authors may feel tempted to add a star to one or two highlighted findings to demonstrate this robustness. Irrespective of their origin, reporting errors undermine the reliability of empirical research and future research may erroneously build on these findings (Azoulay et al., 2015).

We analyze reporting errors in 30,993 tests from 370 articles published in the *American Economic Review* (AER), *Quarterly Journal of Economics* (QJE) and *Journal of Political Economy* (JPE). Our sample mainly comprises hypothesis-testing regression coefficients that address the research question(s) of the respective article and they are typically reported with their respective standard errors and eye-catchers to denote the respective levels of statistical significance. We use an algorithm that flags tests as potential reporting errors and gives authors the benefit of the doubt if rounding may be responsible for apparent reporting errors (e.g., Bruns et al., 2019; Nuijten et al., 2016). We verify the flagged tests by contacting all authors of the afflicted studies. As some flagged tests are not verified due to non-responses by the authors, we draw a random sample of these tests and replicate the corresponding studies. Insights from the replications allow us to further verify flagged tests and to ultimately obtain a reliable estimate of the rate of reporting errors.

Most research on reporting errors has been conducted in psychology. However, statistical reporting in psychology differs from economics. Main findings are usually reported by providing the value of the test statistic accompanied with the degrees of freedom and a p-value, often following the guidelines of the *American Psychological Association* (American Psychological Association, 2010). Most studies that analyze reporting errors in psychology focus on Student's t, ANOVA's F and χ^2 tests and some studies explicitly exclude test statistics from regression analyses and model fitting (e.g., Bakker and Wicherts, 2011; Wicherts et al., 2011). Reporting errors are then diagnosed if the reported p-value differs from the p-value that can be calculated based on the reported test statistic and the corresponding degrees of freedom. For psychology, the share of articles with at least one reporting error is found to vary between 34.9% and 63.0% (Bakker and Wicherts, 2011; Wicherts et al., 2011; Caperos and Pardo, 2013; Bakker and Wicherts, 2014; Veldkamp et al., 2014; Nuijten et al., 2016). At the test level, the rate of reporting errors varies between 4.3% and 12.8%. For *strong* reporting errors for which either the reported p-value or the calculated p-value signals statistical significance but the respective other one does not, the error rates are between 6.3% and 20.5% at the article and between 0.8% and 2.3% at the test level.

Reporting errors have been also analyzed in other fields. For medicine, Garcia-Berthou and Alcaraz (2004) find reporting errors in 31.5% of the analyzed articles and 11.5% of the analyzed tests, based on 44 medical articles published in *Nature* and the *British Medical Journal*. For psychiatry, Berle and Starcevic (2007) find error rates of 36.5% and 14.3% at the level of articles and tests, respectively.[1] For experimental philosophy, these error rates are 38.7% and 6.3% (Colombo et al., 2018) and for innovation research 45.0% and 4.0% (Bruns et al., 2019).

For the statistical reporting style in economics, reporting errors are either characterized by an eye-catcher overstating the significance level compared to the calculated p-value or by an eye-catcher understating the significance level compared to the calculated p-value. The reason for the reporting error may be either in the eye-catcher or in the reported statistical values that are used to obtain the calculated p-value. As readers often glance at eye-catchers and empirical research largely focuses on rejecting null hypotheses, reporting errors with overstated significance levels are more consistent with the incentives in academic publishing while there are usually little incentives in favor of understated significance levels. Comparing the probabilities of reporting errors with overstated and understated significance levels helps to assess whether reporting errors are biased in favor of statistically significant findings. Colombo et al. (2018), Nuijten et al. (2016) and Bakker and Wicherts (2011), for example, find an excess of strong reporting errors with overstated significance levels.

[1] Note that Berle and Starcevic (2007) report an error rate of 10.1% at the article level, but this error rate is obtained by considering all analyzed articles, even those without a single test. See also Table 8.

We shed light on potential sources and predictors of reporting errors by using survey responses and exploratory regression analysis. Previous research from psychology indicates that reporting errors may be related to copy-paste mistakes (Bakker and Wicherts, 2011), that authors of studies with reporting errors are reluctant to share data (Wicherts et al., 2011), that sharing data among co-authors does not seem to be related to the prevalence of reporting errors (Veldkamp et al., 2014) and that outlier removal also does not seem to be related to the prevalence of reporting errors (Bakker and Wicherts, 2014).

The contribution of this article is threefold: (1) We estimate the prevalence of (strong) reporting errors in top economics journals, (2) we assess whether reporting errors are biased in favor of statistically significant findings and (3) we shed light on potential sources and predictors of reporting errors.

Our results show that 35.8% of the analyzed articles contain at least one reporting error corresponding to an error rate of 1.3% at the test level. Strong reporting errors occur in 21.6% of the articles corresponding to 0.5% of all tests. While the error rates at the test level are small, we find that the prevalence of reporting errors with overstated significance levels tend to exceed the prevalence of reporting errors with understated significance levels, indicating a bias towards statistically significant findings. According to the survey responses, most reporting errors stem from errors in the eye-catchers and have their origin in the manual transfer of results from statistical software to word-processing software. Exploratory regression analysis suggests that the availability of software code may be associated with a lower probability of reporting errors.

2 Data and empirical strategy

Data

Analyzing inconsistencies between calculated p-values and reported significance levels requires on the one hand sufficient statistical information to calculate a p-value (e.g. a coefficient with a standard error or t-value) and on the other hand an eye-catcher assigning a specific level of statistical significance (e.g. stars or bold printing). Between 2005 and 2011, 370 empirical articles were published in the AER, the JPE and the QJE that satisfy these two conditions corresponding to 30,993 tests. These tests are exclusively extracted from tables and address the research question(s) of the respective article. Tests routinely conducted, for example, for control variables or descriptive statistics are not considered. The 30,993 tests stem from the comprehensive data of Brodeur et al. (2016) and exclude tests without eye-catchers or insufficient information to calculate p-values.[2]

We extend the data by adding the reported significance level for each test. Usually significance levels are indicated

[2] The original dataset and its description can be downloaded from `https://www.aeaweb.org/articles?id=10.1257/app.20150044`.

Table 1: Distribution of reported statistical values

	AER	JPE	QJE	Total
Tests reported with coef. and se	12247	4685	11235	28167
Tests reported with t/z-statistic	553	246	876	1675
Tests reported with p-value	447	66	638	1151
	13247	4997	12749	30993

Notes: *American Economic Review* (AER), *Journal of Political Economy* (JPE), *Quarterly Journal of Economics* (QJE).

by stars and the table notes clarify how the number of stars relates to different significance levels. We also added information on all significance levels used in the respective table. For example, a table may use the 0.01, 0.05 and 0.1 levels of statistical significance.

Table 1 presents descriptive statistics on our sample. Most tests are reported by providing coefficients with standard errors while t- or z-values and p-values are rarely reported. The largest share of tests stems from articles published in the AER and the QJE, while only about 5,000 tests were extracted from articles published in the JPE.

Empirical strategy

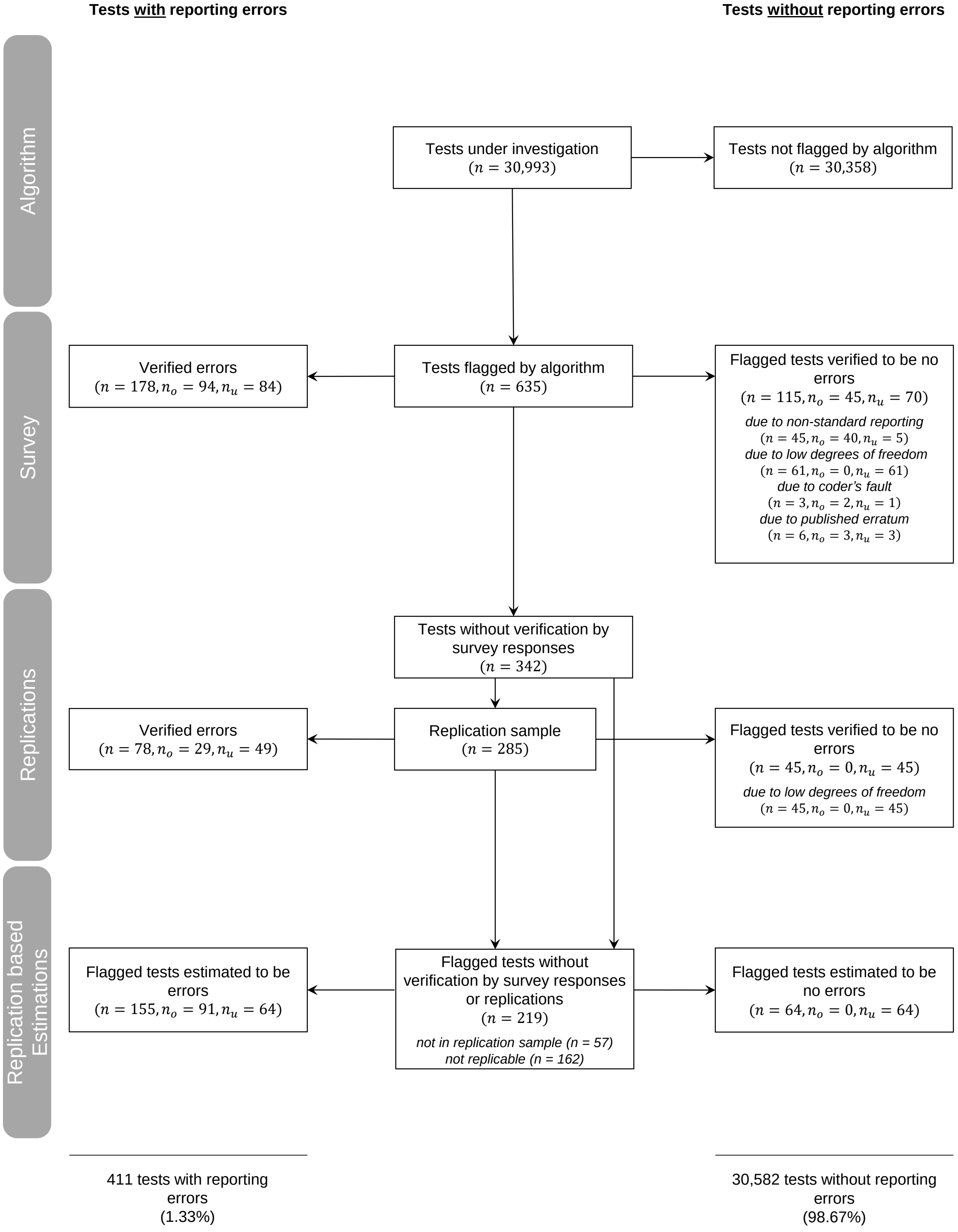
As a first step, we apply an algorithm to all 30,993 tests to flag tests as potential reporting errors (Section 3). This algorithm gives authors the benefit of the doubt by taking the low precision of reported statistical values into account (Bruns et al., 2019; Nuijten et al., 2016). However, some tests may be falsely flagged as potential reporting errors as will be discussed at the end of Section 3. Therefore, we try to validate all flagged tests by sending a survey to the authors of articles with flagged tests (Section 4). As some flagged test remain unverified to be reporting errors due to non-responses by the authors, we draw a random sample of these unverified flagged tests and replicate the underlying articles (Section 5). Based on the replication results, we estimate the rate of reporting errors for all flagged tests that were neither verified by survey responses nor by replications. The flow diagram in Figure 1 visualizes this empirical strategy for the analysis of reporting errors at the level of hypothesis tests. This flow diagram provides also detailed information on the numbers of verified flagged tests and falsely flagged tests for each step of the analysis.

3 Flagging potential reporting errors

De-rounding reported statistical values

The numbers presented in the articles are usually rounded and reported with low precision. In order to account for rounding uncertainties, we calculate intervals consistent with the reported numbers. For instance, a rounded coefficient estimate of 0.019 and a rounded standard error estimate of 0.010 may have their origin in non-rounded estimates from

Figure 1: Empirical strategy to estimate the prevalence of reporting errors at the level of hypothesis tests


Tests with reporting errors
Tests without reporting errors
Algorithm
Survey
Replications
Replication based Estimations
Tests under investigation ($n = 30{,}993$)
Tests not flagged by algorithm ($n = 30{,}358$)
Tests flagged by algorithm ($n = 635$)
Verified errors ($n = 178, n_o = 94, n_u = 84$)
Flagged tests verified to be no errors ($n = 115, n_o = 45, n_u = 70$)
due to non-standard reporting ($n = 45, n_o = 40, n_u = 5$)
due to low degrees of freedom ($n = 61, n_o = 0, n_u = 61$)
due to coder's fault ($n = 3, n_o = 2, n_u = 1$)
due to published erratum ($n = 6, n_o = 3, n_u = 3$)
Tests without verification by survey responses ($n = 342$)
Replication sample ($n = 285$)
Verified errors ($n = 78, n_o = 29, n_u = 49$)
Flagged tests verified to be no errors ($n = 45, n_o = 0, n_u = 45$)
due to low degrees of freedom ($n = 45, n_o = 0, n_u = 45$)
Flagged tests without verification by survey responses or replications ($n = 219$)
not in replication sample (n = 57)
not replicable (n = 162)
Flagged tests estimated to be errors ($n = 155, n_o = 91, n_u = 64$)
Flagged tests estimated to be no errors ($n = 64, n_o = 0, n_u = 64$)
411 tests with reporting errors (1.33%)
30,582 tests without reporting errors (98.67%)

Notes: The flow diagram shows all steps to reach the final estimate of the prevalence of reporting errors at the level of hypothesis tests. n denotes the respective number of tests, n_o the number of tests with overstated significance levels and n_u the number of tests with understated significance levels.

the intervals $[0.0185;0.0194\bar{9}]$ and $[0.0095;0.0104\bar{9}]$. The corresponding possible ratios, that we denote as t-values, are then given by the (rounded) interval $[1.7619;2.0526]$. Generally, reconstructing the degrees of freedom used for the respective test is difficult and often impossible.[3] Therefore, we use critical values from the standard normal distribution rather than from the t-distribution. The interval of possible p-values is then given by $[0.0401;0.0781]$ for a two-sided test with the null hypothesis that the coefficient is equal to zero.[4] Of course, critical values of the t-distribution and standard normal distribution may differ when the degrees of freedom are small. Implications for our analysis are discussed at the end of this section.

For the majority of tests (90.9%), only coefficients and standard errors are reported, but when a test statistic and no p-value is reported (in 5.4% of the tests), the test statistic is also transformed into a p-value interval by taking the low precision of the reported test statistic into account. When a p-value is reported (in 3.7% of the tests), an interval consistent with this potentially rounded p-value is computed. Therefore, we obtain p-value intervals that are consistent with the reported statistical values for all tests in our dataset.

Diagnosis of potential reporting errors

Potential reporting errors are diagnosed if the interval of p-values calculated based on the reported statistical values does not overlap with the interval of p-values assigned by eye-catchers. For example, the coefficient of 0.019 with the standard error of 0.010 implies the p-value interval $[0.0401;0.0781]$. In this case, values on both sides of the threshold of 0.05 are consistent with the reported values and no reporting error is diagnosed if significance at the 0.05 level is stated, giving authors the benefit of the doubt. The p-value interval for the reported statistical significance is obtained by using information from the table notes. For example, if a table reports to use the 0.01, 0.05 and 0.1 levels, a coefficient labelled to be significant at the 0.05 level corresponds to a p-value interval of $[0.01;0.05]$. If the table uses only the 0.05 and 0.1 levels, then the corresponding p-value interval is $[0;0.05]$. An illustration of exemplary reporting errors is given in Table 2.

Prevalence of potential reporting errors

The share of articles with at least one flagged test and the share of flagged tests can be found in the first column in Table 3. In 50.27% (186 of 370) of the analyzed articles, our algorithm flagged at least one test. At the test level, 2.05% of all tests were flagged as being inconsistently reported, corresponding to 635 tests (376 with understated significance levels and 259 with overstated significance levels). 30% (111 of 370) of the articles were flagged as containing at

[3] Sample size is often an insufficient proxy for the degrees of freedom as clustered standard errors are frequently used in the analyzed articles and *Stata* uses the number of clusters as base for the degrees of freedom.

[4] For one-sided tests the p-value interval changes accordingly. Our algorithm accounts for one-sided tests.

Table 2: Exemplary reporting errors

ID	Coefficient	Standard Error	Lower derounding bound of t-value	Upper derounding bound of t-value	p-value interval as implied by reported statistical values	p-value interval as reported by means of eye-catchers	Type of reporting error
1	0.167	0.128	1.2957	1.3147	$0.1890 < p < 0.1951$	$0.05 < p < 0.1$	overstated
2	0.126	0.067	1.8593	1.9023	$0.0571 < p < 0.0630$	$0.01 < p < 0.05$	overstated
3	0.192	0.115	1.6580	1.6812	$0.0927 < p < 0.0973$	$0.1 < p < 1$	understated

Notes: Coefficients and standard errors as reported in the articles. We use two-sided tests and the standard normal distribution to transform the t-value interval into a p-value interval. The eye-catcher and the table notes also imply a p-value interval. If the lower bound of the p-value interval consistent with the reported statistical values is larger than the upper bound of the p-value interval as implied by the eye-catchers, the reported significance level is overstated compared to the calculated p-value. If the upper bound of the p-value interval consistent with the reported statistical values is smaller than the lower bound of the p-value interval as implied by the eye-catchers, the reported significance level is understated compared to the calculated p-value. Bounds rounded to four decimal places.

least one potentially strong reporting error corresponding to 0.69% at the test level (215 tests, among them 109 with understated and 106 with overstated significance levels). While tests with understated significance levels exceed tests with overstated significance levels at both the article and test level, these imbalances become less pronounced for strong reporting errors. Column two and three of Table 3 present refined estimates of the prevalence of reporting errors and are discussed in the next sections.

Limitations of the algorithm

A critical step in our procedure is to treat t-values as being standard normally distributed. The critical values from a t-distribution are greater than their analogues from the z-distribution, especially if the degrees of freedom are small. As a result, the number of flagged tests with understated significance levels may be inflated. For example, if the t-statistic is equal to two and the test is labelled to be only significant at the 0.1 level but the 0.05 level is also used in the respective table, a reporting error with understated significance level seems to be present as two exceeds the critical value of the standard normal distribution for the 0.05 level (1.9600). However, the critical value of the 0.05 level for a t-distribution with, for example, only 50 degrees of freedom is 2.0151 and the reported significance level would actually be correct. The third example in Table 2 illustrates a test that may be falsely flagged as reporting error with understated significance level if the underlying degrees of freedom are small. For example, the critical t-value for 20 degrees of freedom and a significance level of 0.1 is 1.72 and in this case there would be no reporting error present.[5]

A second limitation of the algorithm is related to the style of reporting. The algorithm compares calculated levels of statistical significance based on reported statistical values with reported levels of statistical significance. In some cases, however, the reported statistical values do not directly relate to the reported significance level. Specifically, for non-linear models, such as probit regressions, marginal effects and the corresponding standard errors may be reported in a

[5] The prevalence of errors with overstated significance levels is only affected if authors intentionally use the z-distribution to obtain significance levels in cases when the appropriate t-distribution would lead to a less significant result.

Table 3: Prevalence of reporting errors

			Flagged	Corrected by survey responses	Corrected by survey responses & replications
Article level	Any error	Overstated	103	98	98
			(27.84%)	(26.49%)	(26.49%)
		Understated	143	123	91
			(38.65%)	(33.24%)	(24.64%)
		Any	186	168	133
			(50.27%)	(45.41%)	(35.81%)
	Strong error	Overstated	58	53	53
			(15.68%)	(14.32%)	(14.32%)
		Understated	70	63	45
			(18.92%)	(17.03%)	(12.09%)
		Any	111	102	80
			(30.00%)	(27.57%)	(21.61%)
Test level	Any error	Overstated	259	214	214
			(0.84%)	(0.69%)	(0.69%)
		Understated	376	306	197
			(1.21%)	(0.99%)	(0.64%)
		Sum	635	520	411
			(2.05%)	(1.68%)	(1.33%)
	Strong error	Overstated	106	81	81
			(0.34%)	(0.26%)	(0.26%)
		Understated	109	93	70
			(0.35%)	(0.30%)	(0.23%)
		Sum	215	174	151
			(0.69%)	(0.56%)	(0.49%)

Notes: Numbers and shares of any and strong reporting errors at the article and test level are given. “Overstated” means overstated significance level compared to calculated *p*-value, “Understated” means understated significance level compared to calculated *p*-value. The estimates are based on our algorithm to flag tests in the raw data (first column), after taking into account the survey responses (second column) and after additionally including the information from the replications (third column). The absolute numbers in the third columns are rounded estimates, see Section 5 for the corresponding estimation strategy.

table while the reported significance levels refer to the original model coefficients. We refer to this type of reporting as ’non-standard’ to emphasize that the reported statistical values usually directly relate to the reported significance level. In the case of non-standard reporting, the reported statistical values and the eye-catchers do not report redundant information and we cannot check for reporting errors.

The limitations of the algorithm are addressed in the next two sections by refining the estimated rates of reporting errors based on survey responses from the authors and replications of a random sample.

4 Survey

Survey questions

We sent a survey via email to all authors whose articles contain at least one flagged test to validate the findings of our algorithm and to shed light on the sources of reporting errors. In our first question, the authors were asked where the reporting error occurred, that is, whether it occurred in the coefficient, standard error, test statistic, *p*-value or eye-catcher. Two further response options were "I don't know" and "There is no reporting error". The second question concerned the source of the potential reporting error. As possible response options, we offered: "Error occurred while transferring results from statistical software to word processing software such as Word or Latex", "Error occurred while updating tables during the research/review process", "Error occurred in typesetting by the publisher and remained undetected in proofreading", "Reporting error is falsely diagnosed due to low degrees of freedom of the corresponding test (algorithm to detect reporting errors relies on critical values of the standard normal distribution)", "I don't know", "Other reason" and "If 'other reason' applies, please specify". We sent one reminder to non-responding authors after three weeks and assured the authors to treat their answers anonymously.[6]

Responses

The survey was responded by 89 of 163 contacted authors (54.6%) corresponding to 100 articles (53.8% of all articles containing at least one flagged test) and 323 flagged tests (50.9% of all flagged tests).[7] Regression analyses explaining the propensity to answer to the survey are given in the Online Appendix. All in all, the models have very low explanatory power and no distinct associations between the regressors and the response probability are found.

Authors replied that 133 or 42.0% of all flagged tests are no reporting errors (Table 4). Most of the remaining 184 flagged tests were confirmed to be errors in the eye-catchers. Among these 184 cases, the incorrect transfer of results from statistical software to word processing software such as *Word* or *LaTeX* ("transfer") was the main explanation for reporting errors (Table 5). This answer was given for 39.1% of the errors, almost four times more often than each of the two other main sources: Updating of tables during the research / review process ("updating") and typesetting by the publisher ("typesetting"). 27.2% of the errors were not explained. Other sources were given for 17.4% of the errors.[8]

[6] The email and an exemplary survey can be found in the Supplementary Material.

[7] Six of these flagged tests were due to a misalignment (wrong formatting) in one article. The author pointed out that an erratum was published. Therefore, we classify these flagged tests as no reporting errors and use the remaining 317 flagged tests as benchmark for the analyses of the survey responses. The updates of the error rates in column two and three of Table 3 treat these six tests also as no reporting errors.

[8] These include answers which were not possible to assign reasonably to the other response categories, e.g. rounding errors.

Table 4: Where is the reporting error? ($n = 317$)

Coefficient	Stand. error	Test statistic	p-value	Eye-catcher	There is no error	I don't know
3	12	0	0	135	133	34
(0.9%)	(3.8%)	(0.0%)	(0.0%)	(42.6%)	(42.0%)	(10.7%)

Table 5: Why is there a reporting error? ($n = 184$)

Transfer	Updating	Typesetting	I don't know	Other reason
72	15	15	50	32
(39.1%)	(8.2%)	(8.2%)	(27.2%)	(17.4%)

Notes: "Transfer" refers to the incorrect transfer of results from statistical software to word processing software such as *Word* or *LaTeX*. "Updating" indicates that an error occurred while updating tables during the research/review process."Typesetting" means that an error occurred in typesetting by the publisher and remained undetected in proofreading.

Classification of flagged tests

We classify a flagged test as reporting error if an error was confirmed by the authors, that is, if they replied that the error occurred at a specific place (e.g. coefficient) or due to a particular reason (e.g. typesetting). We cross-checked the 133 flagged tests which the authors replied to be no reporting errors. As can be seen in Table 6, in 21.8% of the cases the authors plausibly argued that low degrees of freedom caused the test to be falsely flagged ("low df"). In other instances, the same reason was given, but we were not able to confirm the argumentation. Most importantly, errors with overstated significance levels cannot be falsely flagged due to low degrees of freedom. We classified those answers as implausible and did the same for other implausible or illogical answers.[9]

A further reason for falsely flagged tests by our algorithm were deviations from the common reporting style in which the reported statistical values directly relate to the reported significance level. This redundant information is needed to check for reporting errors. On the contrary, if a non-linear model is used (e.g., probit model), authors sometimes report the coefficients and standard errors of marginal effects, but the eye-catchers refer to the statistical significance of the original probit coefficients.[10] Non-standard reporting is the reason for 45 falsely flagged tests stemming from five articles with one article accounting for 26 of these tests. Although the answers are plausible to us after validation, a distinct explanation of the reporting style is missing in four of the five articles.

If the authors argued that there was no reporting error but without reasoning, we examined whether erroneous coding, low degrees of freedom, or a non-standard reporting style could have been the reason for falsely flagged tests. We found that for 32 cases low degrees of freedom are a possible explanation and agreed with the authors' responses ("Low df possible"). Data was falsely coded for three flagged tests.

[9] For example, some authors argued that they interpreted significance levels as less than or equal to some value instead of strictly less. However, the probability to obtain a p-value exactly equal to a threshold is zero and it is more likely that in fact a rounding error or another type of error occurred.

[10] An anonymous reviewer pointed out that the well-known *Stata* option *eform* also implies a non-standard reporting style. But this option was not mentioned by the authors in their survey responses.

Table 6: Why is there no reporting error? ($n = 133$)

Coder's fault	Non-standard reporting	Low df	Low df possible	Implausible answer
3	45	29	32	24
(2.3%)	(33.8%)	(21.8%)	(24.1%)	(18.0%)

Notes: "Coder's fault" refers to an error in the original coding or by us. "Non-standard reporting" means that the reported significance level and the reported statistical values are not related. "Low df" refers to low degrees of freedom which cause a falsely flagged test since our algorithm to detect reporting errors relies on critical values of the standard normal distribution. "Low df possible" means that the authors did not give a reason why there is no reporting error, but we found that low degrees of freedom are a likely reason that there is indeed no reporting error. "Implausible answer" indicates that the answer of the author why there should not be a reporting error is implausible.

Update of error rates

We update the rates of reporting errors given in column one of Table 3 by using the survey responses that clarified that some tests were falsely flagged as reporting errors and the updated estimates can be found in column two of Table 3 (Figure 1 visualizes this update as well). In sum, 109 of the initially flagged tests are likely to be no errors with the main reasons being low degrees of freedom (61) and non-standard reporting style (45), see Table 6. Of the 259 tests initially flagged as errors with overstated significance levels, 36.3% were confirmed to be indeed errors, 16.2% were falsely flagged as errors and 47.5% remain without verification from the authors either because the authors did not reply to the survey or replied "I do not know" to both survey questions, see Tables 4 and 5. The 16.2% of tests that were falsely flagged correspond to 42 tests of which 40 used a non-standard reporting style and two were incorrectly coded. As becomes evident in column two of Table 3, the rate of errors with overstated significance levels decreases at the test level moderately from 0.84% to 0.69% for all errors and 0.34% to 0.26% for strong errors while the prevalence at the article level decreases only slightly from 27.84% to 26.49% for all errors and 15.68% to 14.32% for strong errors. The error rate at the article level remains similar as only a few articles account for many falsely flagged tests due to non-standard reporting.

Of the 376 tests initially flagged as understated significance levels, 22.3% were confirmed to be indeed errors, 17.8% were falsely flagged as errors and 59.8% remain without verification. The 17.8% tests that were falsely flagged correspond to 67 tests of which 61 were flagged because of low degrees of freedom, five due to a non-standard-reporting style and one due to a coding error. We expect the number of falsely flagged tests to be higher for tests with understated significance levels due to the limitations of the algorithm. The error rate at the test level moderately reduces from 1.21% to 0.99% for all errors and 0.35% to 0.30% for strong errors while at the article level the prevalence decreases from 38.65% to 33.24% for all errors and 18.92% to 17.03% for strong errors (Table 3, column two). Again, reduction at the article level is smaller as articles often have multiple flagged tests of which not all result from low degrees of freedom.

5 Replications

Replication strategy

The survey sheds light on 293 (46.1%) of the flagged tests and leaves 342 (53.9%) of the flagged tests without manual verification by the authors (Figure 1 provides an overview).[11] In the following, we estimate how many of the flagged tests without verification by the authors are indeed reporting errors by replicating afflicted studies. We took a random sample of 30% from all flagged tests without verification resulting in 103 tests from 63 articles. As we tried to replicate all flagged tests of these 63 articles the sample comprises 285 tests corresponding to 83.3% of all flagged tests without verification. For the calculation of the shares on the test level as presented below respective weights are taken into account. We searched the web for data and software code for the respective articles and used *Stata* 12.1 and *R* 3.5.1 (Windows) to conduct the replications.

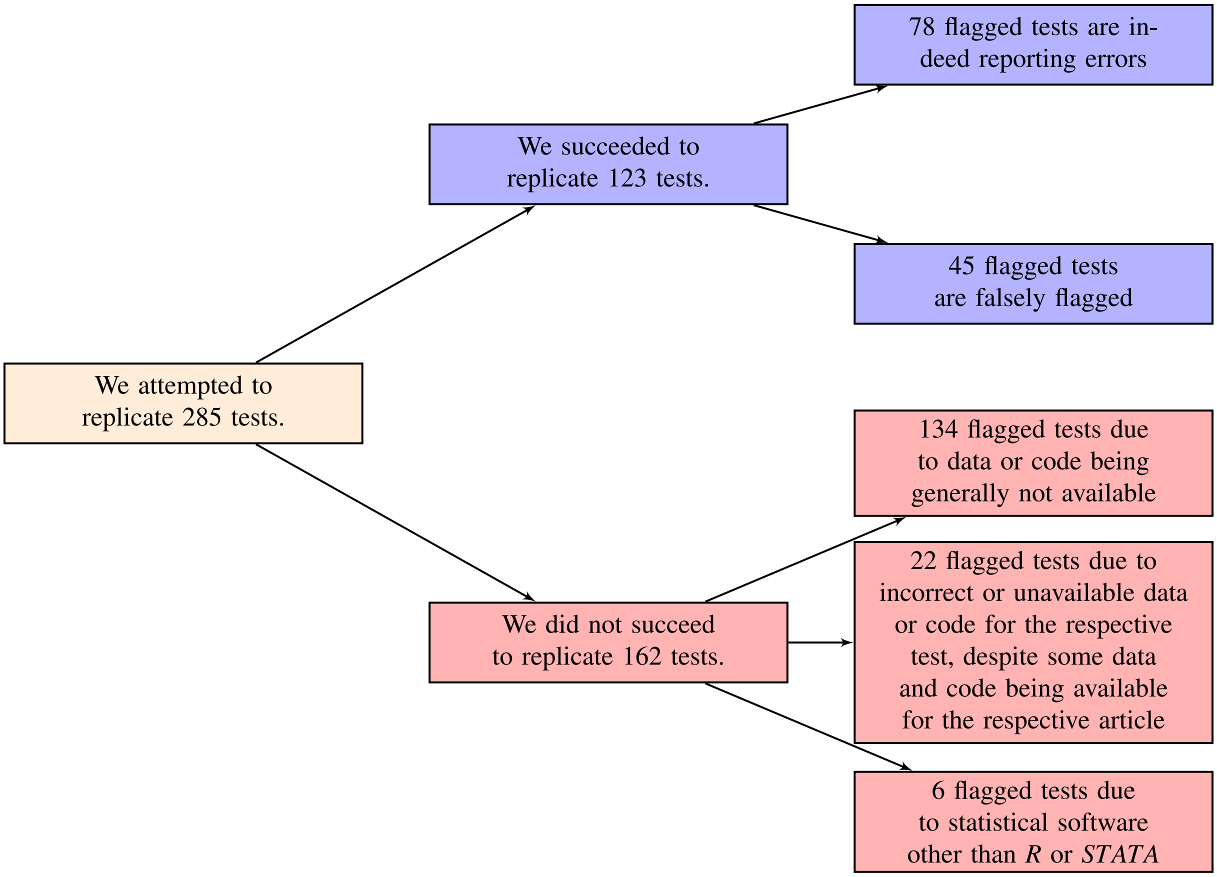
If we were able to replicate the reported statistical values of the flagged test exactly, we checked whether the *p*-value obtained in the replication is consistent with the *p*-value interval reported in the article by means of an eye-catcher. In this case, we classified the flagged test as no reporting error. Sometimes the replicated statistical values were different to those reported in the article, but we succeeded to identify the statistical tests of interest in the software output. In these cases, we used the values reported in the article and calculated the corresponding *p*-value by using the procedure given by the authors' code (degrees of freedom and distribution under the null hypothesis). If this *p*-value was consistent with the eye-catcher reported in the article, we again classified the flagged tests as no reporting errors. This procedure allows us to give the benefit of the doubt to the authors if, for example, software was updated and the same command produces slightly different standard errors today. If data and software code were available but the replication attempt failed, that is, we were not able to identify the respective statistical test in the generated software output, we classified the test as being not replicable but not as a reporting error. In this study, we define reporting errors as inconsistencies between reported levels of statistical significance and calculated *p*-values based on reported statistical information.

Findings

Figure 2 provides an overview of the replication results, while more detailed information on each test in the replication sample can be found in the Supplementary Material. We succeeded to replicate 123 (29 overstated and 94 understated) or 43.2% out of the 285 flagged tests belonging to 22 of the 63 studies (34.9%), among them seven articles containing both replicable and non-replicable tests. For 162 or 56.8% of the 285 flagged tests, we were not able to replicate the results. These tests belong to 48 out of 63 articles (76.2%). The main reason for non-replicability was that data or

[11] 91.2% are due to no response from the authors and 8.8% due to the authors' reply "I do not know" to both survey questions.

Figure 2: Replication results


78 flagged tests are indeed reporting errors
We succeeded to replicate 123 tests.
45 flagged tests are falsely flagged
We attempted to replicate 285 tests.
134 flagged tests due to data or code being generally not available
22 flagged tests due to incorrect or unavailable data or code for the respective test, despite some data and code being available for the respective article
We did not succeed to replicate 162 tests.
6 flagged tests due to statistical software other than *R* or *STATA*

software code was not provided. This was the case for 134 tests (82.7% of 162) from 37 articles (77.1% of 48).[12] In 22 cases (13.6% of 162) from ten articles (20.8% of 48), the replication attempt failed despite data and *Stata* or *R* code being generally available for the corresponding articles, but the code for the flagged test was missing or not working.[13] For the remaining six cases (3.7% of 162) from two articles (4.2% of 48) we did not replicate the tests since a different software was used.

Based on the 123 replicated tests, 100% of the flagged tests with overstated significance levels can be confirmed to be indeed reporting errors.[14] For tests with understated significance levels, the corresponding weighted rate is 50.2%. All falsely flagged tests with understated significance levels occurred due to low degrees of freedom. Regarding strong reporting errors, the weighted rates of correctly flagged strong errors are 100% for errors with overstated significance levels and 55.7% for errors with understated significance levels.

To compute the rate of articles with at least one correctly flagged (strong) error, we divided the number of articles with

[12] Data confidentiality was the reason why data was not available for eight tests from three articles.

[13] In three cases, the software code for the flagged test was available, but generated an error message. In another case, a variable the code referred to was not available in the dataset.

[14] The reliability of this high rate is substantiated in the Online Appendix.

at least one flagged test of a particular kind (overstated, understated or any error) after the replications were conducted by its analog before the replications. The resulting rates of articles with at least one correctly flagged error are 100% (overstated), 54.7% (understated) and 60.2% (any error). The similarity of the latter two shares is explained by the fact that all articles with incorrectly flagged tests with understated significance levels had no other unverified flagged test. The shares of articles with at least one correctly flagged strong error are calculated to be 100% (overstated), 45.2% (understated) and 58.0% (any error). These shares are fairly similar to the ones at the test level as presented above, but based on smaller sample sizes: There are only eight articles with at least one test with overstated significance level, 20 with at least one test with understated significance level and 22 with at least one flagged test of any kind before the replications were conducted. For articles with strong reporting errors, these numbers reduce to five, eleven and fourteen.

Update of error rates

The sum of the flagged tests with overstated and understated significance levels which were neither verified by the authors nor replicated was multiplied by the respective shares of correctly detected errors as presented above. The same was done for the sum of articles with at least one of such non-verified tests, for tests with overstated and understated significance levels and any flagged test, respectively. Following this strategy, we find that in 35.81% of the investigated articles, there is at least one reporting error (Table 3, column three). At the test level, 1.33% of all tests are afflicted by a reporting error. For strong reporting errors these numbers reduce to 21.61% and 0.49% (Table 3, column three). Overall, the rates of tests with overstated significance levels exceed slightly those of tests with understated significance levels. Robustness checks for the estimated error rates can be found in the Online Appendix.

6 Exploratory regression analyses

Model specification

In addition to the survey responses, we explore potential predictors of reporting errors applying logistic regression models. The dependent variable indicates whether an article includes at least one (strong) reporting error or not. We implement the corrections obtained from the survey responses and replications. We run logistic regressions at the article level to avoid the high influence of outliers on the estimates that may occur in an analysis at the test level.[15] Since we did not specify hypotheses beforehand, our analyses should be deemed purely exploratory.

The explanatory variables are taken from the large set of variables gathered by Brodeur et al. (2016) and we focus on

[15] The number of (strong) errors per article has a heavily skewed distribution as is shown in the Online Appendix.

Table 7: Exploratory regression results

	Any error	Any error	Strong error	Strong error
Intercept	71.441	63.117	79.513	59.712
	[-134.315; 273.273]	[-143.469; 270.990]	[-145.139; 316.866]	[-176.435; 305.876]
Year	-0.036	-0.031	-0.040	-0.030
	[-0.137; 0.066]	[-0.135; 0.071]	[-0.158; 0.072]	[-0.152; 0.088]
Journal of Political Economy	0.516	0.074	0.743	0.082
	[-0.052; 1.087]	[-0.650; 0.756]	[0.078; 1.333]	[-0.693; 0.864]
Quarterly Journal of Economics	-0.055	-0.939	-0.142	-1.435
	[-0.528; 0.419]	[-1.813; -0.050]	[-0.711; 0.416]	[-2.351; -0.449]
Field: Macroeconomics	0.624	0.636	0.521	0.562
	[0.123; 1.103]	[0.115; 1.119]	[-0.051; 1.008]	[-0.020; 1.043]
No. of authors	-0.170	-0.181	-0.101	-0.115
	[-0.404; 0.069]	[-0.425; 0.060]	[-0.400; 0.182]	[-0.427; 0.183]
Share of editors among authors	-0.326	-0.370	-0.477	-0.550
	[-0.957; 0.374]	[-1.010; 0.337]	[-1.177; 0.270]	[-1.253; 0.203]
Share of tenured authors	0.643	0.816	0.335	0.591
	[-0.268; 1.492]	[-0.123; 1.694]	[-0.692; 1.277]	[-0.479; 1.524]
Authors' average years since PhD	-0.007	-0.018	0.010	-0.007
	[-0.054; 0.036]	[-0.067; 0.025]	[-0.041; 0.056]	[-0.058; 0.041]
No. of research assistants thanked	-0.046	-0.044	-0.046	-0.044
	[-0.117; 0.021]	[-0.114; 0.024]	[-0.156; 0.038]	[-0.154; 0.041]
No. of individuals thanked	0.013	0.013	0.015	0.014
	[-0.017; 0.039]	[-0.018; 0.040]	[-0.019; 0.045]	[-0.021; 0.045]
Negative results put forward	-0.199	-0.200	-0.161	-0.157
	[-0.769; 0.365]	[-0.785; 0.365]	[-0.865; 0.455]	[-0.896; 0.486]
With theoretical model	-0.417	-0.399	-0.590	-0.594
	[-0.862; 0.030]	[-0.849; 0.083]	[-1.116; -0.067]	[-1.139; -0.045]
No. of tables	0.089	0.105	-0.033	-0.009
	[-0.022; 0.196]	[-0.007; 0.211]	[-0.141; 0.086]	[-0.126; 0.117]
No. of tests	0.005	0.005	0.006	0.007
	[0.002; 0.008]	[0.002; 0.008]	[0.003; 0.010]	[0.003; 0.010]
Data available		0.066		-0.131
		[-0.601; 0.840]		[-0.913; 0.851]
Code available		-1.014		-1.291
		[-1.809; -0.137]		[-2.180; -0.285]
n	367	367	367	367
Pseudo R^2	0.0795	0.0908	0.0782	0.1019

Notes: Results from logistic regressions are shown. The dependent variable is whether an article contains at least one (strong) reporting error or not. Lower and upper bounds of 90% bias corrected and accelerated (BCa) intervals based on 5000 bootstrap replicates in brackets.

those that vary at the article level. In particular, we include the journal, the research field, whether negative results are put forward, whether a theoretical model is used, data availability, code availability, the year of publication, the authors' average years since their PhDs as well as the numbers of authors, research assistants, individuals thanked, tables and tests and the shares of editors and tenured authors among the authors. More details on the variables and descriptive statistics are given in the Online Appendix. We reran the models 5000 times using a nonparametric bootstrap.

Results

The results in the first column and third column of Table 7 are very similar, that is, the probabilities to observe an article with at least one reporting error and an article with at least one strong reporting error can be explained by the same variables. In most of the 5000 bootstrap samples, articles without theoretical models, from the field of macroeconomics in comparison to microeconomics and with more tests are more likely to include at least one (strong) reporting error.[16] Likewise, articles in the JPE seem to be afflicted by at least one (strong) reporting error more frequently than in the AER and QJE. One of the reasons might be the journal policy that mainly determines whether data and software code are published. In our regression sample, in none of the articles in the QJE data or code are available on the website of the journal. The articles in the AER are mostly accompanied by data (82.6%) and code (92.1%), while for the JPE data and code are available in 46.7% and 48.3% of the articles, respectively.[17]

If data and code availability are added to the regression, code availability and the QJE have a negative effect on the probability of at least one (strong) reporting error while the JPE cannot be distinguished from the baseline (AER) over most of the bootstrap replicates (second and fourth column in Table 7). The Pearson correlation between the availability of data and code is 0.79. If code and data are added separately to the model, the effect of code availability is negative with the confidence interval not overlapping zero and the effect of data availability is negative with the confidence interval overlapping zero (results not reported). The described effects are estimated with considerable uncertainty though. Moreover, the effects of most explanatory variables under consideration vary substantially from negative to positive values over 90% of the bootstrap samples. The explanatory power of the models is limited as can be seen by the low Pseudo R^2.

7 Discussion

Our analyses show that reporting errors are present in top economics journals. The estimated error rate of 35.8% at the article level is comparable to what was found in medicine, psychiatry and experimental philosophy and is at the lower end of what was found in psychology. The share of articles with at least one strong reporting error is comparably high with 21.6% while the error rates at the test levels are comparably small with 1.3% for all errors and 0.5% for strong errors. However, comparisons across fields are challenging. Table 8 provides an overview of the previous research on reporting errors, indicating differences in research designs. For instance, the key study on reporting errors in psychology uses the *R* package *statcheck* (Epskamp and Nuijten, 2015) to automatically extract statistical values reported in the text according to the guidelines of the *American Psychological Association* (APA) while tables are

[16] The reference categories for nominal variables are given in the Online Appendix.

[17] While all three journals enforce stringent transparency policies nowadays, there was no change regarding the journal policies in the time frame of our investigation.

excluded from the analysis (Nuijten et al., 2016). The focus on hypothesis tests that are reported in the text and the exclusion of hypothesis tests reported in tables may explain a part of the difference in the rates of reporting errors. Moreover, Nuijten et al. (2016) and Bakker and Wicherts (2014) show that the research design based on *statcheck* tends to detect a higher prevalence of reporting errors compared to manually coded hypothesis tests.

The style of reporting may also explain a part of the difference in the rates of reporting errors. In economics, the key findings are usually presented as regression coefficients in tables with corresponding standard errors and eye-catchers to denote the levels of statistical significance. In the other fields summarized in Table 8, main findings are usually reported by providing the value of the test statistic accompanied with the degrees of freedom and a p-value.

Our analysis is most comparable with Bruns et al. (2019) who analyze a sample of 5,667 hypothesis tests from 101 articles published in *Research Policy*. Our algorithm is based on Bruns et al. (2019) and they also consider regressions coefficients reported in tables with either standard errors, t-values, z-values or p-values and eye-catchers that denote the levels of statistical significance. They find 45.0% of the articles to have at least one reporting error corresponding to 4.0% at the level of hypothesis tests. For strong reporting errors, the rates are 25.0% and 1.4% at the level of articles and hypothesis tests, respectively. These rates of reporting errors are larger than those found in the present study for top economics journals. Bakker and Wicherts (2011) find that the rate of reporting errors is smaller in three high impact psychology journals compared to three psychology journals with lower impact. While the most prestigious group of journals in economics is usually considered to be the "top 5" (and all three journals analyzed here belong to this group), the impact factor of *Research Policy* (5.351) is, according to the *Journal Citation Report 2019* (Clarivate Analytics, 2020), only slightly smaller than those of the AER (5.561) and the JPE (5.504) but much smaller than the one of the QJE (11.375). Hence, more research is needed on whether and how the prevalence of reporting errors in economics is associated to the journal's impact and prestige.

We compare the probability of a strong reporting error with overstated significance level with the probability of a strong reporting error with understated significance level to assess whether a bias in favor of statistically significant estimates is present. Understated significance levels are usually neither favorable nor consistent with the incentive system in academic publishing, since statistically significant results as signaled by eye-catchers are usually considered to be of most interest to the readers. We find 81 (0.26%) strong reporting errors with overstated significance levels and 70 (0.23%) strong reporting errors with understated significance levels, which indicates a small imbalance towards strong reporting errors with overstated significance levels. Based on these rates, it is 15.71% more likely to find a strong reporting error being reported as statistically significant than to find a strong reporting error being reported as non-significant.

However, comparing the probabilities of strong reporting errors with overstated and understated significance levels

Table 8: Prevalence of reporting errors in the current study and related studies

Article	Field	Journal and article selection criteria	Test selection criteria	Position of tests	Data collection	Error rate at article level	Error rate at article level (strong)[a]	Error rate at test level	Error rate at test level (strong)[a]	Number of articles including at least one investigated test	Number of investigated tests
Current article	Economics	Three top economics journals	Tests that address the research questions of the article: mostly regression coefficients with either standard error, t-value, z-value or p-value and an eye-catcher	Table	Manually	35.8	21.6	1.3	0.5	370	30,993
Bruns et al. (2019)	Innovation research (economics and management)	*Research Policy*	Tests that address the research questions of the article: mostly regression coefficients with either standard error, t-value, z-value or p-value and an eye-catcher	Table	Manually	45.0	25.0	4.0	1.4	101	5,667
Nuijten et al. (2016)	Psychology	Eight top psychology journals	t, F, r, χ^2 and Z statistics reported according to APA guidelines	Text	*statcheck*	49.6	12.9	9.7	1.4	16,695	258,105
Veldkamp et al. (2014)	Psychology	Six top psychology journals	t, F, r, χ^2 and Z statistics reported according to APA guidelines	Text	*statcheck*	63.0	20.5	10.6	0.8	430	8,105
Bakker and Wicherts (2014)	Psychology	Articles including the word “outlier“ from main psychology journals	t and F tests with degrees of freedom and $p < .05$	Text and table	*statcheck* and manually	45.1	15.0	6.7	1.1	153	2,667
Caperos and Pardo (2013)	Psychology	Four Spanish psychology journals	Student's t, ANOVA's F and χ^2 tests with degrees of freedom and p-value apart from regression analysis and model fitting	Text and table	Manually	48.0	17.6	12.2	2.3	102	1,212
Wicherts et al. (2011)	Psychology	Two main psychology journals	t, F and χ^2 test statistics with degrees of freedom and $p < .05$ apart from model fitting	Text and table	Manually	53.1	14.3	4.3	0.9	49	1,148
Bakker and Wicherts (2011) (Study 1)	Psychology	Three high-impact and three low-impact psychology journals	Student's t, ANOVA's F and χ^2 tests with degrees of freedom and p-value apart from regression analysis and model fitting	Text and table	Manually	54.6	18.0	9.7	1.2	194	4,077
Bakker and Wicherts (2011) (Study 2)	Psychology	Random sample from all type of psychology journals	Student's t, ANOVA's F and χ^2 tests with degrees of freedom and p-value apart from regression analysis and model fitting	Text and table	Manually	34.9	6.3	12.8	1.1	63	643
Colombo et al. (2018)	Experimental Philosophy	Sample mostly from philosophy journals	t, F, r, χ^2 and Z statistics reported according to APA guidelines	Table	*statcheck*	38.7	6.4	6.3	0.5	173	2,573
Berle and Starcevic (2007)	Psychiatry	Two psychiatry journals	Student's t, ANOVA's F and χ^2 tests with with degrees of freedom and exactly reported p-value	Text and table	Manually	36.5	9.4	14.3	-	96[b]	546
Garcia-Berthou and Alcaraz (2004)	Medicine	Medical papers from *Nature* and *British Medical Journal*	All test statistics with degrees of freedom and a precise p-value	Text and table	Manually	31.5	-	11.5	0.4	44	244

[a] As opposed to our study, all other studies except for Caperos and Pardo (2013) who are not clear in this respect consider a fixed p-value threshold to define strong errors. Bruns et al. (2019) base their calculations of strong errors on $p = 0.1$, otherwise $p = 0.05$ is considered.

[b] Berle and Starcevic (2007) investigate tests with exactly reported p-values, but only report the number of articles including any kind of significance statement. Thus, the 96 articles are an upper bound for the number of articles including at least one investigated test.

in all tests is agnostic with respect to the cause of the strong reporting errors. Survey responses indicate that the vast majority of errors might be due to a misreporting of eye-catchers (see Section 4). This is also consistent with the survey responses in Bruns et al. (2019). If strong reporting errors appear exclusively due to a misreporting of eye-catchers, then the probability of a strong reporting error with overstated significance level is given by the conditional probability that a truly non-significant test is misreported as being statistically significant. These truly non-significant tests are the tests reported to be non-significant plus the strong reporting errors with overstated significance levels (as these are actually non-significant) minus the strong reporting errors with understated significance levels (as these tests are actually statistically significant). In our sample, the share of strong reporting errors with overstated significance levels in the truly non-significant tests is 0.64%. The probability of a strong reporting error with understated significance level is given by the conditional probability that a truly significant test is misreported as being non-significant. These truly significant tests are the tests reported to be significant plus the strong reporting errors with understated significance levels (as these are actually significant) and minus the strong reporting errors with overstated significance levels (as these tests are actually non-significant). In our sample, the share of strong reporting errors with understated significance levels among the truly significant tests is 0.38%.[18] Hence, if we assume that all strong reporting errors are due to misreported eye-catchers, it would be 66.47% more likely to find a truly non-significant test being misreported as statistically significant than to find a truly significant test being misreported as non-significant.

These numbers indicate a tendency for a bias in favor of statistically significant estimates. Such a bias was also found in the field of psychology (Nuijten et al., 2016; Bakker and Wicherts, 2014) and the field of experimental philosophy (Colombo et al., 2018). In innovation research, Bruns et al. (2019) focus on the analysis at the article level and conclude that there is little indication of a bias in favour of statistically significant findings as the imbalance between articles with at least one reporting errors with overstated significance level and articles with at least one reporting errors with understated significance level evens out if strong reporting errors are considered.[19] However, their findings at the level of tests are more mixed as they depend on the inclusion or exclusion of three studies that may be considered as outliers due to the large number of errors per article.

While the error rates found in our study tend to be comparably small, our analysis indicates that reporting errors might be biased in favor of statistically significant findings. It is important to stress that there are many reasons for this tendency. Authors are likely to be more willing to accept a statistically significant estimate than an estimate that is inconsistent with their prior beliefs (Bastardi et al., 2011; Kunda, 1990) and, thus, many reporting errors may be the result of human errors and biases rather than intentional misbehavior. Moreover, the publication process is generally

[18] When only considering the flagged tests as in the first column of Table 3, the share of strong reporting errors with overstated significance levels in the truly non-significant tests is 0.83% and the share of strong reporting errors with underrstated significance levels in the truly significant tests is 0.60% . After accounting for the survey responses (second column of Table 3), these values change to 0.64% and 0.51% , respectively.

[19] In our sample, it is is 17.78% more likely to observe an article with at least one strong reporting error with overstated significance level compared to an article with at least one strong reporting error with an understated significance level.

biased towards statistically significant estimates and this is presumably also true for tests that are only seemingly significant due to a reporting error. Moreover, questionable research practices cannot be excluded as a potential source. In psychology, rounding down *p*-values to let them appear statistically significant was found to be a common misbehavior (John et al., 2012). In economics, the equivalent of this misbehavior might be that authors add stars to marginally non-significant estimates. But to the best of our knowledge there is no evidence on how common such a research practice is in economics and our analysis cannot shed light on this.

With regard to the sources of reporting errors, the survey respondents named the manual transfer of results from statistical software to word processing software as the major source and replied that errors predominantly occurred in eye-catchers. Even though automatic procedures had existed a long time before 2005 when our time frame of investigated articles starts, e.g. *outreg* for *Stata* (Gallup, 1998), it is conceivable that a manual transfer might still have been common practice in those days and it might still be today. Bakker and Wicherts (2011) find for psychology that the reporting of one finding was frequently used as a template for the reporting of other findings and that this copy-pasting resulted in errors.

Our exploratory regression analyses show that the prediction of reporting errors is difficult with the help of the available variables. One reason is measurement error in the dependent variable as our algorithm to detect reporting errors is not perfect. Most apparent is the negative association between the availability of software code and (strong) reporting errors. It has to be noted that code availability is highly correlated with data availability and that the availability of both differs substantially between the three journals. In psychology, Wicherts et al. (2011) find that authors of studies with reporting errors are reluctant to share data while Veldkamp et al. (2014) find no association between data sharing among co-authors during the research process and the prevalence of reporting errors. More research is needed to understand how data and code availability are associated with the presence of reporting errors and how this is linked to journal policies regarding the availability of data and software code.

8 Conclusions and recommendations

We use an algorithm to flag potential reporting errors in more than 30,000 hypothesis tests from 370 articles published in the *American Economic Review*, *Journal of Political Economy* and *Quarterly Journal of Economics*. We refine the estimated rate of reporting errors by sending a survey to the authors of all studies with flagged tests and by replicating a large random sample of studies. Our final estimates suggest that 1.3% of all hypothesis tests are afflicted by a reporting error while 0.5% of all tests are afflicted by a strong reporting error. This relates to 35.8% of the articles being afflicted by at least one reporting error while 21.6% are afflicted by at least one strong reporting error. In other words, every fifth article contains at least one reporting error for which either the eye-catcher or the calculated *p*-value (based on the reported statistical values, such as coefficients and standard errors) signals statistical significance but the respective other one does not. These errors might influence how readers perceive the conclusiveness or robustness of the findings reported in a respective article.

We also find a bias in favor of strong reporting errors with overstated significance levels, that is, eye-catchers signal statistical significance but the reported statistical values do not imply statistical significance. While this type of error is consistent with the incentive system in academic publishing, it remains unclear to what extent human errors and biases or intentional misbehavior contribute to these errors.

We find at least weak indication that software code availability is associated with a lower probability of reporting errors. Survey respondents name the manual transfer of results from statistical software to word processing software as the major source of reporting errors and replied that errors are mostly due to misreported eye-catchers. All of these findings are exploratory and can be used to generate hypotheses for future research.

We conclude with four recommendations that may help to reduce the rate of reporting errors in future research:

First, journals should oblige researchers to make their data and code available. Authors that provide data and code are likely to carefully check whether the uploaded code indeed reproduces the published tables and this is likely to reduce the probability of a reporting error. More importantly, transparency facilitates replications and permits others to check the accuracy of the published findings.[20] Similar arguments have been made, among others, by Chang and Li (2017).

Second, building on the first recommendation, a more vivid replication culture should be incentivized by introducing replication sections in top journals (Coffman et al., 2017) and by accepting positive replications for publication (Mueller-Langer et al., 2019).

Third, eye-catchers to denote statistical significance should be banned. According to the authors who participated in our survey, most of the detected reporting errors stem from errors in the eye-catchers. Moreover, eye-catchers facilitate the bad practice to judge scientific results in a binary way according to arbitrary thresholds. It is conceivable that eye-catchers distract authors and reviewers from studying intensively the actual statistical values and checking them for consistency. The *American Economic Association* indeed nowadays forbids authors to use stars to denote statistical significance in their journals, among others in the *American Economic Review*.[21] However, eye-catchers are still omnipresent in other journals.

Fourth, errors can be reduced in the research, review and publication process by simple measures. Automatic software should be used to transfer statistical results to word processing software. Algorithms such as the one used in this paper could be used in the review process and by the authors themselves after the typesetting as suggested by Nuijten et al.

[20] As pointed out by an anonymous reviewer, one further option to enhance transparency and to guarantee exact replicability is to mirror and archive the computer that was used for the estimation procedure.

[21] `https://www.aeaweb.org/journals/aer/submissions/accepted-articles/styleguide`

(2016).

In principle, the four recommendations above are easy to implement, yet they require a cultural shift in economics. For instance, refraining from old standards such as attaching stars to statistical results or publishing well-documented software code may be burdensome for journal editors and researchers, respectively. New regulations and incentives may reinforce better habits but also need to be sustained and accepted. Fortunately, the attention for the reliability of empirical research has increased and better research practices in economics are on the rise (Christensen et al., 2019).

References

Albarqouni, L. N., J. A. López-López, and J. P. Higgins (2017). Indirect evidence of reporting biases was found in a survey of medical research studies. *Journal of Clinical Epidemiology 83*, 57–64.

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

Azoulay, P., J. L. Furman, J. L. Krieger, and F. Murray (2015). Retractions. *Review of Economics and Statistics 97*(5), 1118–1136.

Bakker, M. and J. M. Wicherts (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods 43*(3), 666–678.

Bakker, M. and J. M. Wicherts (2014). Outlier removal and the relation with reporting errors and quality of psychological research. *PLoS ONE 9*(7), 1–9.

Bastardi, A., E. L. Uhlmann, and L. Ross (2011). Wishful thinking: Belief, desire, and the motivated evaluation of scientific evidence. *Psychological science 22*(6), 731–732.

Berle, D. and V. Starcevic (2007). Inconsistencies between reported test statistics and p-values in two psychiatry journals. *International Journal of Methods in Psychiatric Research 16*(4), 202–207.

Brodeur, A., M. Lé, M. Sangnier, and Y. Zylberberg (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics 8*(1), 1–32.

Bruns, S. B., I. Asanov, R. Bode, M. Dunger, C. Funk, S. M. Hassan, J. Hauschildt, D. Heinisch, K. Kempa, J. König, J. Lips, M. Verbeck, E. Wolfschütz, and G. Buenstorf (2019). Errors and biases in reported significance levels: Evidence from innovation research. *Research Policy 48*(9), 103796.

Bruns, S. B. and J. P. Ioannidis (2016). P-curve and p-hacking in observational research. *PloS one 11*(2), e0149144.

Caperos, J. M. and A. Pardo (2013). Consistency errors in p-values reported in Spanish psychology journals. *Psicothema 25*(3), 408–414.

Chang, A. C. and P. Li (2017). A preanalysis plan to replicate sixty economics research papers that worked half of the time. *American Economic Review: Papers & Proceedings 107*(5), 60–64.

Christensen, G., Z. Wang, E. L. Paluck, N. Swanson, D. J. Birke, E. Miguel, and R. Littman (2019). Open science practices are on the rise: The state of social science (3s) survey. *UC Berkeley CEGA Working Paper*. Available from `https://escholarship.org/uc/item/0hx0207r`.

Clarivate Analytics (2020). Journal impact factor, journal citation reports, 2019.

Coffman, L. C., M. Niederle, and A. J. Wilson (2017). A proposal to organize and promote replications. *American Economic Review 107*(5), 41–45.

Colombo, M., G. Duev, M. B. Nuijten, and J. Sprenger (2018). Statistical reporting inconsistencies in experimental philosophy. *PloS one 13*(4), e0194360.

Doucouliagos, H., M. Paldam, and T. D. Stanley (2018). Skating on thin evidence: Implications for public policy. *European journal of political economy 54*, 16–25.

Epskamp, S. and M. B. Nuijten (2015). statcheck: Extract statistics from articles and recompute p values. *R package version 1.3.0*. Available from `http://CRAN.R-project.org/package=statcheck`.

Franco, A., N. Malhotra, and G. Simonovits (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science 345*(6203), 1502–1595.

Gallup, J. (1998). Formatting regression output for published tables. *Stata Technical Bulletin 8*, 28–30.

Garcia-Berthou, E. and C. Alcaraz (2004). Incongruence between test statistics and p-values in medical papers. *BMC Medical Research Methodology 4*(1), 13.

Gerber, A. and N. Malhotra (2008a). Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Quarterly Journal of Political Science 3*(3), 313–326.

Gerber, A. S. and N. Malhotra (2008b). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research 37*(1), 3–30.

Ioannidis, J. P., T. D. Stanley, and H. Doucouliagos (2017). The power of bias in economics research. *Economic Journal 127*(605), F236–F265.

John, L. K., G. Loewenstein, and D. Prelec (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science 23*(5), 524–532.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review 2*(3), 196–217.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin 108*(3), 480–498.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review 73*(1), 31–43.

Mueller-Langer, F., B. Fecher, D. Harhoff, and G. G. Wagner (2019). Replication studies in economics – how many and which papers are chosen for replication, and why? *Research Policy 48*(1), 62–83.

Munafò, M. R., B. A. Nosek, D. V. Bishop, K. S. Button, C. D. Chambers, N. Percie Du Sert, U. Simonsohn, E. J. Wagenmakers, J. J. Ware, and J. P. Ioannidis (2017). A manifesto for reproducible science. *Nature Human Behaviour 1*(1), 1–9.

Nuijten, M. B., C. H. Hartgerink, M. A. van Assen, S. Epskamp, and J. M. Wicherts (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods 48*(4), 1205–1226.

O'Boyle, E. H., G. C. Banks, and E. Gonzalez-Mulé (2017). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management 43*(2), 376–399.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin 86*(3), 638–641.

Simmons, J. P., L. D. Nelson, and U. Simonsohn (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science 22*(11), 1359–1366.

Veldkamp, C. L., M. B. Nuijten, L. Dominguez-Alvarez, M. A. Van Assen, and J. M. Wicherts (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PLoS ONE 9*(12), 1–19.

Vivalt, E. (2019). Specification searching and significance inflation across time, methods and disciplines. *Oxford Bulletin of Economics and Statistics 81*(4), 797–816.

Wasserstein, R. L., N. A. Lazar, et al. (2016). The asa's statement on p-values: context, process, and purpose. *The American Statistician 70*(2), 129–133.

Wicherts, J. M., M. Bakker, and D. Molenaar (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE 6*(11), 1–7.

Supplementary Material

The Supplementary Material contains the Online Appendix, an exemplary survey email, more detailed replication results as well as data and code. The Supplementary Material is available at `https://github.com/peterpuetz2020/reporting_errors_economics`.