Statistical reporting errors in economics

Non Peer-reviewed author version

BRUNS, Stephan; Herwartz, Helmut; Ioannidis, John P.A.; ISLAM, Chris-Gabriel &
Raters, Fabian H. C. (2023) Statistical reporting errors in economics.

DOI: 10.31222/osf.io/mbx62
Handle: http://hdl.handle.net/1942/42667

Statistical reporting errors in economics*

Stephan B. Bruns[1,2,3], Helmut Herwartz[1], John P. A. Ioannidis[3], Chris-Gabriel Islam[1,2,4,†] and Fabian H. C. Raters[1]

[1]Georg August University of Göttingen, Germany
[2]Hasselt University, Belgium
[3]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, USA
[4]Federal Statistical Office of Germany
[†]Corresponding author: chris-gabriel.islam@uni-goettingen.de

Abstract

We developed a tool that scrapes and interprets statistical values (DORIS) to analyze reporting errors, which occur if the eye-catcher depicting the level of statistical significance is inconsistent with the reported statistical values. Using 578,132 tests from the top 50 economics journals, we find that 14.88 % of the articles have at least

*Acknowledgements: We thank Igor Asanov, Guido Bünstorf, Chris Doucouliagos, Thomas Kneib, Peter Pütz, Tom Stanley, Katharina Zigova, Nina Baueregger and the team of the Göttingen State and University Library, and participants at the MAER-Net conferences 2021 in Athens and 2022 in Tokyo, Meta-Research Day 2019 in Tilburg, VfS Annual conference 2022 in Basel and seminar and summer/fall school presentations at Georg August University of Göttingen in 2020 and 2021. We are grateful to Niklas Döbbeling, Leo Dörr and Jens Lichter for excellent research assistance. The views expressed in this article are solely those of the authors and do not necessarily reflect those of the Federal Statistical Office of Germany. Stephan Bruns, Helmut Herwartz and Chris-Gabriel Islam are grateful for funding from the German Research Foundation (DFG) under the project "Replications in Empirical Economics: Necessity, Incentives and Impact" with follow up project "Selective reporting and the evolving research landscape in economics" (project number: 405039391).

We currently work on providing DORIS at betterpapers.org as a service for authors, reviewers, and editors to easily check their papers for reporting errors.

one strong error in the main tests. Our pre-registered analysis suggests that mandatory data and code availability policies reduce the prevalence of strong errors, while suggestive indication of a reversed effect is found for top 5 journals. Integrating DORIS into the review process can help improving article quality.

Keywords: Reporting errors; Replications; Reproducibility; Data and code policies; Questionable research practices
JEL codes: A19, C18, C40, C87

Main Contributions

- First large-scale assessment of reporting errors in economics based on 578,132 statistical tests from the top 50 economics journals.
- We developed a tool that automatically scrapes and interprets statistical values from tables in economics journals (DORIS - Diagnosis Of Reporting errors In Scraped tables), potentially useful for improving the review process and future meta-research.
- We substantially improve accuracy of flagging reporting errors by estimating the degrees of freedom.
- We show that reporting errors are prevalent in economics and exhibit a systematic bias in favor of significant results.
- Suggestive evidence is found that mandatory data and code availability policies reduce the prevalence of strong reporting errors.
- Our results also provide suggestive indication that the top 5 journals have a larger probability of strong reporting errors.

1 Introduction

Statistical reporting errors received scant attention in economics compared with other disciplines such as psychology (e.g., Nuijten et al. 2016; Wicherts et al. 2011; Caperos and Pardo 2013; Veldkamp et al. 2014) and medicine (García-Berthou and Alcaraz 2004; Berle and Starcevic 2007). A typical and important type of reporting error in economics occurs if a regression coefficient is highlighted to be statistically significant by means of an eye-catcher (mostly stars or asterisks), but actually the reported coefficient and standard error do not imply the presence of such a level of statistical significance. Such an inconsistency might invalidate conclusions made in an article and mislead the reader. We define reporting errors as an inconsistency between reported levels of statistical significance using eye-catchers and calculated p-values implied by reported statistical values, such as coefficients and standard errors. We focus our analysis on *strong* reporting errors for which either the reported level of statistical significance or the calculated p-value signals statistical significance at the chosen level of the authors but the respective other one of these two does not. This is opposed to *weak* reporting errors which occur if the inference on whether a statistical test is statistically significant or not remains unchanged (cf. Section 2.2.1).[1]

In psychology, Nuijten et al. (2016) developed an automated procedure to scrape statistical tests that are reported in the text according to the American Psychological Association (APA) guidelines (*statcheck*, see Epskamp and Nuijten 2018), allowing the analysis of reporting errors in more than 250,000 tests. Such an automated analysis of reporting errors is more challenging in economics as statistical tests are reported heterogeneously and predominantly in tables. In this article, we use recent big data tools like web scraping and text mining and develop an automated procedure to reliably scrape and interpret statistical values from economic articles. Our algorithm is called DORIS (**D**iagnosis **O**f **R**eporting errors **I**n **S**craped tables). DORIS is able to download published articles in HTML and extract statistical tests from tables as well as additional information from the text.

[1] Some authors would define p-values below 0.05 as significant while few authors even suggest p-values below 0.2 as significant. The usual threshold in economics is 0.1.

DORIS collected data from 31 journals between 1998 and 2016.[2] In total, our sample comprises 3,746 articles with 578,132 statistical tests extracted from tables. We use these data to assess the prevalence of reporting errors in economics on a large scale and to analyze potential determinants.

There is little research on statistical reporting errors in economics. While Bruns et al. (2019) found for a sample of about 6,000 tests from *Research Policy* (RP) that 1.4 % of the tests contain a strong reporting error and 25 % of the articles are afflicted by at least one strong reporting error, Pütz and Bruns (2021) found a prevalence of 0.5 % at the test level and 21.6 % at the article level in a study of about 30,000 tests from *American Economic Review* (AER), *Journal of Political Economy* (JPE) and *Quarterly Journal of Economics* (QJE) which all belong to the top 5 economic journals. Both studies rely on manually collected statistical tests and are thus limited in the range of considered journals, time spans, and the number of analyzed articles and tests. Error rates found in other disciplines tend to be similar, but reporting styles of statistical tests differ substantially, e.g., key statistical findings in psychology are generally reported in the text following specific reporting guidelines by the APA. For psychology, Nuijten et al. (2016) found a prevalence of 1.4 % strong reporting errors at the test level and 12.9 % at the article level. Other studies in psychology found strong reporting error rates between 0.8 % and 2.3 % at the test level and between 6.3 % and 20.5 % at the article level. In medicine, García-Berthou and Alcaraz (2004) found a strong reporting error rate of 0.4 % at the test level. In psychiatry, Berle and Starcevic (2007) found a strong reporting error rate of 9.4 % at the article level. Lastly, in experimental philosophy, Colombo et al. (2018) found a prevalence of 0.5 % strong reporting errors at the test level and 6.4 % at the article level. Table B.3 in the Online Appendix provides an overview and further details.

The source of these reporting errors might be questionable research practices (QRP) like rounding down a p-value such that it appears statistically significant (e.g., Wicherts et al. 2011; John et al. 2011) or it might just be an honest mistake, introduced by manually transferring the output of statistical software to word processing software or at the journal typeset-

[2]Due to the dispute between Elsevier and German universities, we stop at 2016 as our sample drastically decreases to only 19 journals from 2017 and beyond.

ting stage (Pütz and Bruns 2021; Bakker and Wicherts 2011). Indication was also found that the prevalence of reporting errors might be negatively associated with the availability of software code (Pütz and Bruns 2021) and the unwillingness to share data (Wicherts et al. 2011), while data sharing among co-authors (Veldkamp et al. 2014) or outlier removal (Bakker and Wicherts 2014) does not appear to have an effect. Although we do not expect conscious fraud to explain most reporting errors, our sample includes articles whose primary conclusions are affected once reporting errors are taken into account.

We can distinguish between *overstated* and *understated* reporting errors. An overstated reporting error occurs if a statistical test is marked by an eye-catcher implying that the reported significance level is *overstated* compared to the p-value calculated based on the reported statistical values. Alternatively, the eye-catcher may imply that the reported significance level is *understated* compared to the calculated p-value. Various studies, as e.g., Colombo et al. (2018), Nuijten et al. (2016), Bakker and Wicherts (2011) and Pütz and Bruns (2021), find a higher prevalence of overstated reporting errors in comparison to understated reporting errors, indicating a bias towards significant findings. This bias might be due to the fact that the significant findings are more in line with the authors' expectations and therefore are not double checked, or due to publication bias towards significant findings (Nuijten et al. 2016). However, not all studies have found such a bias in favor of overstated errors (Weinerová et al. 2022).

Moreover, we shed light on potential covariates of reporting errors using a pre-analysis plan that we uploaded to the Open Science Framework (OSF)[3] to clearly distinguish between hypothesis-testing and exploratory research (e.g., Olken 2015; Christensen and Miguel 2018). With the first set of pre-registered hypotheses, we try to fill a gap revealed by Christensen and Miguel (2018) who mention that open data and code are assumed to improve research quality. Recent examples show how data and code availability policies help reveal already published mistakes (Bach et al. 2023; Matray and Boissel 2023), but quantitative meta-analytical evidence is missing. We hypothesize that mandatory data and code availability

[3]Bruns et al. (2021, January 11). Statistical reporting errors in economics. Open Science Framework (OSF), https://osf.io/tqmuj, last retrieved on 25.08.2023.

policies are associated with the probability of a test being afflicted with a strong overstated or a strong understated reporting error. Recently, an increasing number of journals require or at least encourage authors to publish their data and code, and such policies might be an effective instrument to reduce the prevalence of reporting errors. Based on previous findings that the prestige and impact of journals might be associated with reporting error rates (Bakker and Wicherts 2011), the second set of pre-registered hypotheses states that being a test published in a top 5 economics journal affects the probability to be afflicted with a strong overstated or a strong understated reporting error. Such an association could be interpreted as indication that the rigor of peer review could affect the prevalence of reporting errors, since the top 5 journals play a special role in economics (Ductor et al. 2020; Card and DellaVigna 2013; Heckman and Moktan 2020) and generally have longer peer review processes (Ellison 2002).

We find a prevalence of strong reporting errors of 0.46 % at the test level and 30.89 % of the articles contain at least one strong reporting error. When considering only tests that are likely to address the main hypotheses of the respective article, we find a prevalence of strong reporting errors of 0.51 % at the test level and of 14.88 % at the article level. 1.37 % (2.04 %) of all tables in our sample have more than 10 % of the tests exhibiting strong reporting errors when considering all tests (only the main tests). This emphasizes the relevance of reporting errors as readers might be misled. Our results also indicate a systematic bias towards statistically significant reporting errors. Pre-registered and exploratory research indicates that open data and code policies reduce the prevalence of strong overstated reporting errors at the test level. The effect of these policies on strong understated reporting errors is ambiguous. We find no evidence that belonging to the top 5 journals changes the probability of strong understated reporting errors, but some indication that the probability of strong overstated reporting errors might be increased.

Our findings suggest that statistical reporting errors are non-negligible in economics, as is the case in other disciplines. Using DORIS in the review process may help authors, reviewers, and editors identify reporting errors before publication and can help to improve article quality. Data and code

availability policies may be decisive in ensuring replicability and are likely to be effective in reducing reporting error rates.

In the next section, we will provide an overview of the development and application of DORIS and our data. In Section 3 we present results on the prevalence and relevance of reporting errors in economics. Then, we present our hypothesis testing analysis in Section 4 including our pre-registered regression analysis as well as an exploratory difference-in-difference design. Sections 5 and 6 discuss and conclude, respectively. For a more detailed description of DORIS and additional analyses, we refer to a comprehensive Online Appendix in the next sections.

2 Diagnosis of reporting errors in scraped tables (DORIS)

2.1 Web scraping and rule-based interpretation

We consider the top 50 economic journals[4] for the web scraping and the rule-based interpretation of statistical values. From these journals, 32 journals offered their articles as HTML files and allowed us to web scrape them (cf. Table B.2).[5] We focus on articles available in HTML, as this ensures a more reliable scraping process compared to PDF documents.[6] We downloaded all articles published between 1998 and 2016. From these 32 journals that are available in HTML, we extracted the HTML tables and then developed a rule-based program to identify and interpret statistical values from these tables. Only three journals provide their articles in HTML before

[4]The top journals are based on the IDEAS/RePEc Ranking Simple Impact Factors (Last 10 Years) as of December 2018: https://ideas.repec.org/top/old/1812/top.journals.simple10.html, last retrieved on 04.11.2022.

[5]Although articles from journals with *Springer* as a publisher (*Journal of Economic Growth*, *IMF Economic Review* and *Experimental Economics*) were also available in HTML, we deemed *Springer*'s conditions to scrape these journals to be not acceptable for a scientific project. We are grateful to the publishers *Oxford University Press*, *The University of Chicago Press*, *Wiley*, *Elsevier*, *Annual Reviews* and *Taylor & Francis* who generously allowed us to scrape their articles for this project.

[6]In HTML documents a table is explicitly embedded in a table environment marked by "<table>" and "</table>", while current PDF extractors like Tabula only work with computer vision techniques and positioning of characters, resulting in unreliable extractions especially for complicated types of tables.

1998 (see Table B.2 in the Online Appendix) and in 2016, the number of accessible journals drops substantially to 19 journals due to the dispute between the publisher *Elsevier* and the German universities.[7] As the p-value interval of one-sided tests is different from the one of two-sided tests and we cannot automatically detect one-sided tests at the test level, we focus on two-sided regression tests and removed all tests from articles that mention the term "one-sided" or a similar expression which excluded 5.8 % from the sample.[8]

For the development of the rule-based approach, a large sample of tables with manually interpreted statistical tests and the corresponding significance levels is needed. We used 360 tables of the data gathered by Brodeur et al. (2016), 258 tables of the data gathered by Bruns et al. (2019), and we have additionally drawn a random sample of 500 tables from the top 5 general-interest journals (from which QJE, JPE and *Review of Economic Studies* (RESTUD) are part of our sample due to their availability in HTML) and seven additional top field journals (i.e. based on the list in Table B.2: *Journal of Finance* (JOF), *Economic Policy* (EP), *Journal of Monetary Economics* (JME), *Journal of Labor Economics* (JOLE), *Journal of Development Economics* (JODE), *Journal of Applied Econometrics* (JOAE) and *Journal of Public Economy* (JPUE)). The rule-based interpretation is based on the following major principles: First, tables with eye-catchers are identified mostly by comparing cell contents with a list of possible eye-catchers; second, different table styles are identified, such as standard errors below coefficients or standard errors next to coefficients; third, the table notes are used to link eye-catchers to significance levels. A detailed summary of the rule-based interpretation is given in Section A.1 of the Online Appendix. The rule-based approach is able to find 99.9 % of all statistical tests and their respective significance levels reported in the 360 tables obtained from Brodeur et al. (2016) and in the 258 tables obtained from Bruns et al. (2019). For the random sample of 500 tables from the top 5 general-interest and top field journals 92.7 % of all statistical tests are

[7]For more information consult e.g., https://www.science.org/news/2016/12/thousands-german-researchers-set-lose-access-elsevier-journals, last retrieved on 26.09.2021.

[8]The exact search term as well as additional regular expressions can be found in Table B.7 in the Online Appendix.

found by DORIS.[9] Finally, we applied DORIS to the top 5 journals and the seven additional top field journals and then randomly selected 100 identified statistical tests per journal stratified at the article level. We further improved the rule-based interpretation based on the mistakes that DORIS made in this last part of the development.

We evaluated the performance by applying DORIS to all 32 journals in the sample and randomly selected 100 identified statistical tests per journal stratified by volumes and with the condition of not selecting two tests from the same article. We evaluated how often DORIS made a mistake in interpreting the statistical tests and the corresponding significance levels. The False Discovery Rate (FDR) of identifying statistical tests was very small with 1.2 % over all journals. In these cases, DORIS misinterpreted the values of a statistical tests (e.g., a t-value was interpreted as a standard error), or it identified a test that is actually not a test at all (e.g., mean and standard deviation in a descriptive table are interpreted as a statistical test). The FDR broken down by journals can be found in Table B.4 in the Online Appendix. A graphical overview of how DORIS was developed and how its performance in interpreting statistical values from tables was evaluated can be found in Figure C.4 in the Online Appendix.

Before manual control and removal of further implausibilities[10] DORIS extracted and interpreted 729,930 tests from 4,613 articles out of 25,583 downloaded articles in HTML format. A large fraction of the articles do not provide statistical tests that can be analyzed by DORIS. First, around 21 % of the articles do not contain any tables (e.g., theoretical articles). Second, there is an increasing tendency to use confidence intervals in recent years. DORIS has not yet been developed to interpret confidence intervals. Third, some authors opt against using eye-catchers. Note that the AER, which is not in our sample as they do not provide articles in HTML, even

[9] We also implemented one version of DORIS with additional rules to reach 100 % but some of these rules were based on inferring missing information for very specific tables (e.g., deriving missing information about the numbers in parentheses by looking at other tables in the same article instead of assuming the default case with standard errors) or allowed more than one test type in a table which was prone to more errors in the interpretation.

[10] Further details can be found in Section 2.2.2.

forbid the usage of eye-catchers in their guidelines.[11] Finally, note that we only focused on tests for which DORIS could extract the amount of observations and for which DORIS estimated the degrees of freedom to be positive.[12].

2.2 Diagnosis of statistical reporting errors

2.2.1 Definition of statistical reporting errors

We consider two distinct dimensions in the diagnosis of statistical reporting errors (e.g., Pütz and Bruns 2021). First, we distinguish between *overstated* and *understated* reporting errors. Overstated reporting errors refer to errors where the eye-catcher implies a smaller p-value than the reported statistical values, and *vice versa* for understated reporting errors. In these cases, we implicitly assume that the reported statistical values are correct and the eye-catcher are incorrect. An illustrative example is given in Table 1. In the first column, an understated reporting error is shown as the t-value of the variable *primary schooling* is 3.78 with degrees of freedom of 3,000 but the coefficient is only labeled significant at the 0.05 level. In the second column, an overstated reporting error is illustrated as the t-value is 1.98 with degrees of freedom of 3,000 but the coefficient is labeled significant at the 0.01 level.

Second, we define a reporting error to be *strong* if either the reported significance level or the calculated p-value signals statistical significance at the chosen level of the respective authors but the respective other does not. The variable *primary schooling* in the third column of Table 1 depicts such a strong reporting error, as the level of significance reported suggests statistical significance but the reported t-value of 0.77 with degrees of freedom of 3,000 does not. We refer to *weak* reporting errors if the inference on whether a statistical test is statistically significant or not remains unchanged. An example is given in the second column of Table 1 as the t-value is 1.98, which is significant at the 0.05 level but not at the 0.01 level as the eye-catcher would suggest.

[11] Cf. https://www.aeaweb.org/journals/aer/submissions/accepted-articles/styleguide, last retrieved on 26.09.2021.

[12] Details on the extraction of the number of observations and on the estimation of the degrees of freedom can be found in Section 2.2.2

Table 1: Illustrative table with hypothetical values for primary schooling in growth regressions

	(1)	(2)	(3)	(4)	(5)
Primary schooling	0.13**	0.13***	0.13**	0.13*	0.13
	(3.78)	(1.98)	(0.77)	(1.70)	(1.65)
(...)					
Degrees of freedom	3,000	3,000	3,000	20	20

Notes: t-values in parentheses and t-tests are two-sided. *p<0.1; **p<0.05; ***p<0.01.

We also consider the category *any* error comprising any type of error regardless of whether it is a weak or strong reporting error or an understated or overstated reporting error.

2.2.2 Identifying statistical reporting errors

Diagnosing statistical reporting errors proceeds in two steps. First, statistical tests are flagged by DORIS as potential reporting errors. The second step comprises a manual check of these flagged errors. Regarding the flagging of potential reporting errors, Bruns et al. (2019) and Pütz and Bruns (2021) assumed t-values to be normally distributed due to missing information regarding degrees of freedom. This procedure led to an underestimation of overstated errors and an overestimation of understated errors. The underestimation can be seen in the fourth column of Table 1. If the t-value of 1.70 is assumed to be normally distributed, the reported significance level of 0.1 is correct. Actually, the degrees of freedom for this t-test are only 20 and the corresponding critical t-value is 1.725, resulting in a strong overstated reporting error. The overestimation is exemplified in the fifth column. If the t-value of 1.65 is assumed to be normally distributed, the reported non-significance with regard to the 0.1 level is incorrect, resulting in a strong understated reporting error. But actually the degrees of freedom for this t-test are again 20 with the corresponding critical t-value of 1.725, resulting in a test without any reporting error.

DORIS substantially improves the flagging procedure by estimating the degrees of freedom. DORIS extracts the highest number of observations

reported in a given table and uses the maximal number of either table rows or table columns as the number of estimated parameters. This results in an interval $I_r(df) = [r_1(df); r_2(df)]$ of possible p-values based on the reported statistical values (e.g., coefficient with standard error), the estimated degrees of freedom and the rounding uncertainty.[13] This interval has to be compared with the interval of possible p-values $I_e = [e_1; e_2]$ indicated by the eye-catcher. A reporting error occurs if $I_r(df)$ and I_e do not intersect. Bruns et al. (2019) and Pütz and Bruns (2021) used the normal distribution instead of the t-distribution and therefore assumed the degrees of freedom as infinity in $I_r(df)$. This is of no concern if the reported information is either a p- or a z-value. In other cases, and as we only are able to roughly estimate the exact degrees of freedom, we calculate two critical thresholds for the degrees of freedom. The first critical threshold refers to understated reporting errors:

$$df_{crit}^{under} = \min\{df \in \mathbb{N} | r_2(df) \leq e_1\}. \tag{1}$$

Here, $\mathbb{N}$ does not contain zero. We flag a test as being afflicted with an understated reporting error if df_{crit}^{under} is defined[14] and if the estimated degrees of freedom are greater than df_{crit}^{under} as t-values need to be greater to show significance if the degrees of freedom are small. Analogously, we define the critical threshold for overstated reporting errors:

$$df_{crit}^{over} = \sup\{df \in \mathbb{N} | e_2 \leq r_1(df)\}. \tag{2}$$

We choose the supremum instead of the maximum in order to guarantee that df_{crit}^{over} can also take the value of infinity which is the degree of freedom in a normal distribution. We flag a test as being afflicted with an overstated reporting error if the estimated degrees of freedom are less than

[13] With regard to rounding uncertainty, take e.g., a reported coefficient estimate of 1.2 with a standard error 0.6. Naively calculated, this results in a t-value of 2, but when controlling for rounding uncertainties, an interval from $1.\overline{769230}$ to $2.\overline{27}$ should be considered as a rounded 2 could be a value between 1.95 and 2.05 and a rounded 0.6 could be a value between 0.55 and 0.65. We declare a test only as afflicted with a reporting error if such an interval does not intersect with the respective interval that belongs to the reported significance level given by the eye-catcher.

[14] df_{crit}^{under} is not defined if $\{df \in \mathbb{N} | r_2(df) \leq e_1\}$ is the empty set.

df_{crit}^{over}.[15] This procedure leads to a smaller overestimation of understated reporting errors and a smaller underestimation of overstated reporting errors in comparison to Bruns et al. (2019) and Pütz and Bruns (2021). A visualization for better understanding can be found in Section A.2. However, the approach of estimating the degrees of freedom does not account for control variables, constants, or fixed-effects that are part of the model but that are not reported explicitly in the table. Hence, a manual control is still needed.

The second step in diagnosing reporting errors is manually checking the flagged tests. As introduced in Section 2.1, the FDR of statistical tests is small with 1.2 %. However, falsely interpreted statistical values could result in an erroneously flagged statistical test. For example, a t-value that is interpreted as a standard error could yield an erroneously flagged test. Hence, a manual inspection of the flagged tests is necessary, as otherwise we might report inflated error rates.

Before the manuaĺ control, DORIS flags 13,969 tests in 1,893 articles as afflicted with a strong reporting error. Every flag was manually checked. The checked data set might slightly underestimate the real prevalence of reporting errors as DORIS might oversee some statistical tests due to uncommon reporting styles.[16] Comparing with Brodeur et al. (2016) yields that we find 28 % to 70 % more tests per article, which is plausible as we consider every test while Brodeur et al. (2016) focused on main tests.

During the manual control, we focus on the core data of a test, as e.g., the eye-catcher, coefficient and standard error. Additionally, we correct the meta-data, such as mentions of robustness checks in the table or the usage of clustered standard errors, and calculate the exact degrees of freedom. As our approach of calculating the degrees of freedom merely using the number of observations and the number of parameters might not be justified in the case of clustered standard errors, we manually collect the number of clusters. We recalculate the degrees of freedom using the number of clusters instead of the number of observations. If this explains the reported signif-

[15] If $\sup\{df \in \mathbb{N} | e_2 \leq r_1(df)\}$ is the empty set, df_{crit}^{over} equals $-\infty$.

[16] Comparing the median of tests per table in the development data and the median of tests per table in *all_ data* (cf. Figure C.8 in the Online Appendix) yields an average difference of -2.75 tests. Hence, the power/coverage of DORIS at the table level is satisfactory.

icance, we remove the flag in order to give the authors the benefit of the doubt. A detailed explanation of the steps during the manual control can be found in Section A.3 in the Online Appendix. After manually checking the flagged reporting errors, we identified 5,316 strong reporting errors out of 1,296 articles. Similarly to the evaluation phase, most commonly DORIS misinterpreted the redundant information or labeled a summary statistic as a statistical test. We decided to exclude additional 139 articles that we deem as outliers. These are mostly articles that contain a large number of reporting errors due to a misreporting of the numbers in brackets or parentheses by the authors. Holmes (2004) coined this as Reporting Imprecision. This means that the second number of a test, next to the coefficient, is not specified correctly or specified at all, e.g., standard errors are marked as t-values. In order to separate this error, which can usually be easily spotted and corrected by the reader from real reporting errors, we excluded these and further outliers from our sample. A detailed explanation of the outlier removal can be found in Section A.5 in the Online Appendix. After removing outlier articles, we identified 2,675 strong reporting errors out of 1,157 articles. This is our primary sample for the subsequent analyzes. Figure C.8 in the Online Appendix gives an overview of the data sets used and how they are compiled.

We distinguish between main tests that analyze the hypothesis of an article and non-main tests using a heuristic approach. We identify main tests in two steps. First, we exclude tables in the appendix and tables that contain robustness checks.[17] Second, we focus on the first three rows as discontinuity tests suggest that selective reporting is most prevalent in the first three rows of a table (cf. Figure C.3 in Section A.6 in the Online Appendix). We also report results exclusively for the first row as a robustness check. As a complementary data set, we define non-main tests as tests that either appear in tables in the appendix or that are robustness checks or tests that appear in regular tables but not in the first three rows. An overview of the different data sets broken down by the types of tests

[17]Check Table B.7 in the Online Appendix for the regular expressions used to identify these tables.

is given in Table 2. Most of the tests collected by DORIS are tests with coefficients and standard errors.[18]

Table 2: Types of tests per data set

	All tests		Main tests		Non-main tests		First row	
	Amount	[%]	Amount	[%]	Amount	[%]	Amount	[%]
Coefficient and standard error	413,543	71.53	129,496	75.27	284,047	69.95	50,165	76.23
t- or z-value	138,807	24.01	37,502	21.80	101,305	24.95	13,840	21.03
p-value	25,782	4.46	5,042	2.93	20,740	5.11	1,805	2.74

Notes: Sometimes tests provide various values, i.e. p-values alongside coefficients and standard errors. Hence, test types presented are in aggregated order, meaning that a test of the type "p-value" might also contain t- or z-values or even standard errors and a test of type "t- or z-value" might contain standard errors, as well.

3 Descriptive analysis

3.1 Prevalence of statistical reporting errors

Error rates for the different types of statistical reporting errors broken down at the levels of journals, articles, tables, and tests are shown in Table 3. DORIS collects all statistical tests from tables, whereas previous literature often focuses on the main tests of a respective article. We report error rates for all tests, as well as for the subsets shown in Table 2.[19] Almost every journal is afflicted by at least one reporting error. Focusing on strong reporting errors, we find an error rate of 30.89 % for all tests at the article level. Other studies focus mainly on the main tests of the respective article. For these, we find an error rate of 14.88 % which is in the same ballpark found by other studies.[20] If we consider only the first row for defining main

[18]In order to determine the type of the redundant information DORIS text-mines the table notes. If no information is given or can be extracted from the table notes, DORIS chooses the major case based on Pütz and Bruns (2021), which is "coefficient with standard error". If the default case results in an error rate of more than 40 % at the test level of the respective table and if the assumption of either a t- or a p-value results in fewer reporting errors, the respective case is chosen.

[19]Robustness checks for the first two, four and five rows can be found in Table B.8 in the Online Appendix. The error rate at the table, article, and journal level rises the more rows we include as more tests are considered. The error rate at the test level decreases marginally as more rows are considered, leading to the suspicion that reporting errors occur more often in the upper rows.

[20]For economics, 21.6 % in Pütz and Bruns (2021) and 25.0 % in Bruns et al. (2019) while including other fields, the ballpark grows to an interval from 6.4 to 25.0 % (cf. Table B.3 in the Online Appendix).

tests, we find a lower prevalence at the article level of 7.21 %. For non-main tests we find a prevalence of 23.95 % at the article level, which is intuitive, since we consider more non-main tests to be part of the article than main tests. While our error rates are at a similar ballpark as found in other disciplines at the article level, our study finds the prevalence at the test level to be one of the lowest for strong reporting errors with 0.46 % for all tests and 0.51 % for the main tests. It should be noted that these error rates constitute lower bounds, as DORIS was programmed in a conservative way that gives authors the benefit of the doubt.

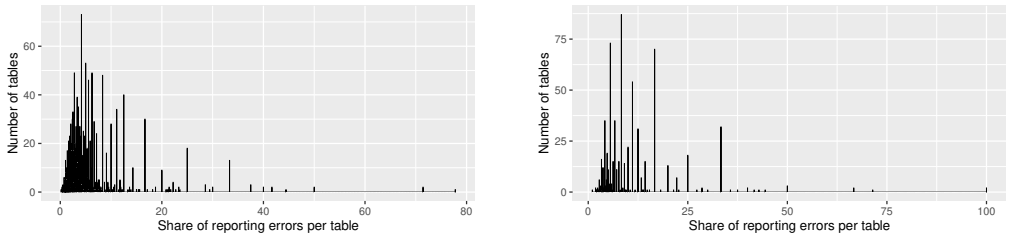
Table 3: Prevalence of statistical reporting errors

		All tests			Main tests			Non-main tests			First row		
Level	Type	Over-stated	Under-stated	Any	Over-stated	Under-stated	Any	Over-stated	Under-stated	Any	Over-stated	Under-stated	Any
Journal	Any	90.32	93.55	93.55	93.33	93.33	93.33	83.87	90.32	90.32	83.33	86.67	90.00
	Strong	83.87	87.10	87.10	86.67	90.00	90.00	74.19	83.87	83.87	60.00	76.67	76.67
Article	Any	30.35	42.66	55.21	16.56	25.75	36.12	22.74	34.28	44.70	8.68	14.99	21.51
	Strong	17.70	19.54	30.89	8.43	7.51	14.88	12.88	15.40	23.95	3.78	3.62	7.21
Table	Any	11.33	18.77	26.70	6.16	10.53	15.81	8.83	15.58	21.73	2.91	5.27	7.95
	Strong	5.67	6.28	11.11	2.79	2.47	5.13	4.41	5.40	9.09	1.10	1.07	2.16
Test	Any	0.54	1.14	1.67	0.67	1.38	2.05	0.48	1.03	1.51	0.74	1.42	2.16
	Strong	0.21	0.25	0.46	0.26	0.24	0.51	0.19	0.26	0.44	0.27	0.26	0.53
No. of tests (articles)		578,132 (3,746)			172,040 (3,677)			406,092 (3,611)			65,810 (3,677)		
No. of tests (articles) afflicted with a strong reporting error		2,675 (1,157)			874 (547)			1,801 (865)			348 (265)		

The error rates as well as the total number of identified tests vary between journals (cf. Figure C.9 and Figure C.10 in the Online Appendix). The overall error rate seems to be declining over time for most journals, which is in line with Nuijten et al. (2016) who find that error rates in psychology have been stable or even declining over time. The overall number of tests increases over time (cf. Figure C.6 in the Online Appendix) as well as the number of tests per article (cf. Figure C.7 in the Online Appendix), indicating a trend in economics toward more analyses per article or more variables included per regression.

3.2 Relevance of statistical reporting errors

Figure 1 shows the distribution of the share of reporting errors among individual articles and tables. This gives indication that reporting errors are not exclusively single appearances in an article or table, but that there are accumulations as well. 215 of 1,747 tables with at least one strong



(a) All tests (b) Main tests

Figure 1: Share of strong reporting errors per table among tables with at least one strong reporting error among all tests

reporting error have an error rate considering all tests larger than 10 %,i.e. 1.37 % of all tables. When we reduce this analysis to main tests, we find that 275 of 692 tables with at least one strong reporting error have an error-rate higher than 10 % in the first three rows, i.e. 2.04 % of all tables that are neither robustness checks nor appear in the appendix.

We further emphasize the relevance of reporting errors by providing anonymized examples in which statements in the article are no longer supported once the reporting error is corrected (Section A.7 of the Online Appendix). Reporting errors in main tests can easily mislead readers and multiple articles have been retracted due to such errors.[21]

3.3 Overstated vs understated reporting errors

Overstated reporting errors are more in line with the incentives in academic publishing, whereas it is generally not desirable to present non-significant findings (e.g., Bruns et al. 2019). There are of course exceptions, e.g., if a previously as significant published finding is contested (Ioannidis 2023). In most circumstances, understated reporting errors may be considered to be more random errors. Therefore, we use understated errors as a baseline to compare with overstated errors to assess whether a bias towards statistically significant findings is present.

Nuijten et al. (2016) compare the number of strong overstated reporting errors in tests labeled significant to the number of strong understated reporting errors in tests that are not labeled significant. These error rates are informative to assess the probability of an overstated reporting error

[21] https://tinyurl.com/mrbv72s4, last retrieved on 02.12.2022.

when looking at the tests that were published as statistically significant and *vice versa* for understated reporting errors.

In contrast, Pütz and Bruns (2021) base their analysis on the rate of overstated reporting errors in truly non-significant findings and the rate of understated reporting errors in truly significant tests. These error rates put more emphasis on how the reporting error occurred. As indicated by the survey in Bruns et al. (2019), reporting errors occur mainly due to misreported eye-catchers. Hence, the probability of a strong overstated reporting error is given by the probability that a truly non-significant test is misreported as being statistically significant. The number of truly non-significant tests can be calculated as follows:

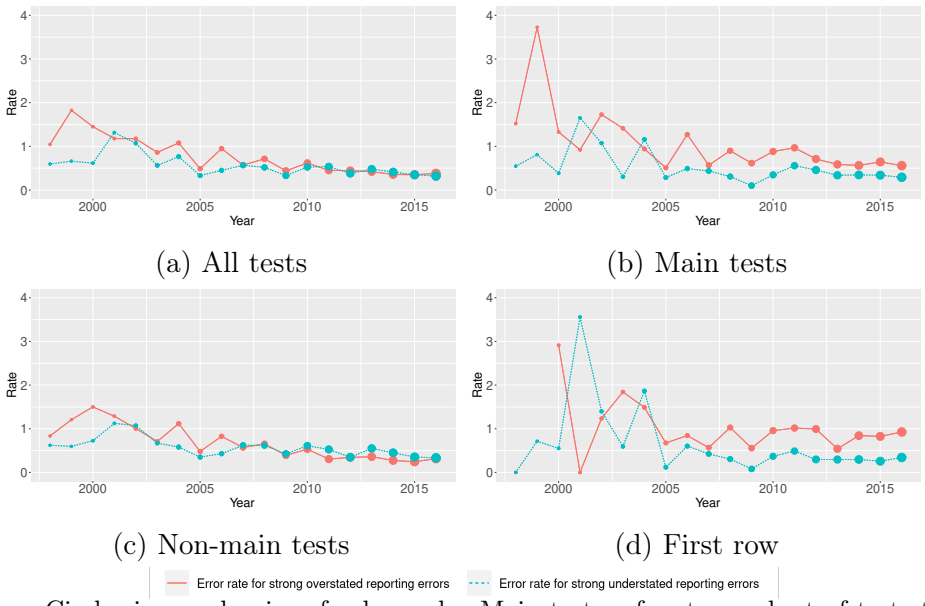
$$\#\text{truly non-significant tests} = \#\text{tests without any eye-catcher} + E_o - E_u, \tag{3}$$

where E_o is the number of tests with a strong overstated reporting error (they are actually non-significant) and E_u is the number of tests with a strong understated reporting error (they are actually statistically significant). Analogously, the number of truly significant tests can be calculated as follows:

$$\#\text{truly significant tests} = \#\text{tests with an eye-catcher} - E_o + E_u. \tag{4}$$

In total, 0.49 % of the truly non-significant tests are strong overstated reporting errors and 0.44 % of the truly significant tests are strong understated reporting errors. Hence, when assuming that all strong reporting errors are due to misreported eye-catchers and in contrast to the numbers in Table 3, there is no suggestive difference between over- and understated reporting errors when we look at all tests. A visualization over the whole time span is depicted in the top left chart of Figure 2. The outliers at the beginning of each of the eight time series can be explained by some specific articles that comprise many reporting errors as can be seen by the size of the small circle which represents the number of tests that are either truly significant or truly non-significant per year. Based on the top left chart, visual inspection does not suggest a systematic bias towards significant tests.

If we focus on the main tests as outlined in Section 2.2.2 and shown in



(a) All tests (b) Main tests

(c) Non-main tests (d) First row

Notes: Circle size marks size of subsample. Main tests refers to a subset of tests that appear in the first three rows of tables that are neither robustness checks nor appear in the appendix. First row refers to a subset of tests that appear in the first row of tables that are neither robustness checks nor appear in the appendix. The larger a dot, the more tests are in the respective sample. Sizes are not comparable across subfigures.

Figure 2: Rate of strong overstated reporting errors among truly non-significant tests and rate of strong understated reporting error errors among truly significant tests

the top right chart of Figure 2, the rate of strong overstated reporting errors exceeds the rate of strong understated reporting errors in most years except early years with a limited number of observations. In total, 0.72 % of the truly non-significant tests are strong overstated reporting errors and 0.39 % of the truly significant tests are strong understated reporting errors. Considering only the first row of tables that are neither robustness checks nor appear in the appendix, 0.87 % of the truly non-significant tests are strong overstated reporting errors and 0.38 % of the truly significant tests are strong understated reporting errors, widening the gap even further up to 0.49 percentage points (cf. bottom right chart of Figure 2). These numbers even exceed the number of Pütz and Bruns (2021) who find an excess of 0.26 percentage points for strong overstated reporting errors. In sum, this analysis suggests that there is a tendency for a systematic bias towards statistically significant reporting errors, especially if we focus on our two definitions of main tests.

4 Hypothesis-testing analysis

4.1 Hypotheses

Concerning the prevalence of strong statistical reporting errors, we formulated four hypotheses in total that we registered at the OSF using a pre-analyis plan:[22]

1. Journal policies regarding the availability of data and software code are associated with the probability that a test is afflicted by (a) a strong overstated reporting error; (b) a strong understated reporting error.

2. Belonging to the 'top 5' journals is associated with the probability that a test is afflicted by (a) a strong overstated reporting error; (b) a strong understated reporting error.

Regarding the effect of data and code availability policies, Pütz and Bruns (2021) provide exploratory evidence that the availability of data and/or code might reduce the prevalence of reporting errors at the article level in economics. Note that Pütz and Bruns (2021) focused on the actual availability of data and/or code per published article while we focus on journal policies.[23] Note that, following the definition of Vlaeminck (2021), we focus on mandatory data availability policies which require authors to upload their replication files to a third party prior publication, while we neglect policies that only encourage authors to upload their files or author responsibility policies which leave the responsibility to share data to the authors. While Nuijten et al. (2017) do not find any association between reporting errors and data availability policies in psychology, Wicherts et al. (2011) show that the unwillingness to share data is associated with the prevalence of reporting errors. Nosek et al. (2021) suggests that open data sharing can reduce or expose reporting errors in the field of psychology. These previous findings indicate that journal policies might be effective in reducing the rate of reporting errors in economics.

[22] Bruns et al. (2021, January 11). Statistical reporting errors in economics. Open Science Framework (OSF), https://osf.io/tqmuj, last retrieved 25.08.2023.

[23] Authors may provide data and code without a journal policy imposing this, and some authors may not provide data and code even if a policy requires them to do so (Christensen et al. 2019, e.g.,).

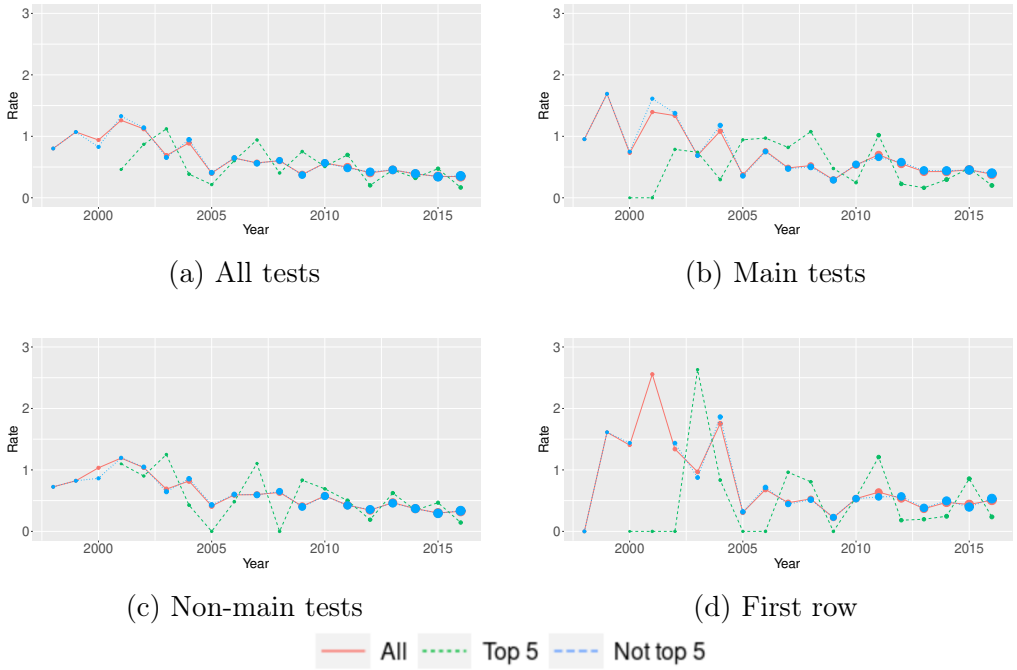
We created a dummy variable to indicate whether the respective test was published in a journal that had an active open data and software code policy in the year of the publication. We deem it necessary that the policy includes both, data and code, as a full replication is difficult if a researcher lacks one of it. Regarding our data set, a total of 120,519 tests belong to this category. Table B.5 in the Online Appendix provides a journal-specific overview.

Regarding the second hypothesis, Bakker and Wicherts (2011) show that the prevalence of reporting errors in high impact psychology journals is less than the prevalence in lower impact psychology journals. The error rates found by Pütz and Bruns (2021) in the QJE, the AER and the JPE, which belong to the top 5 economic journals, are smaller than the error rates found by Bruns et al. (2019) for RP which does not belong to the IDEAS/RePEc list of top 50 economic journals, although it has a similar journal rank based on the 2019 Journal Citation Report (Clarivate Analytics 2020). These prior findings indicate that the journal's prestige might be associated with the rate of reporting errors.

We coded a dummy variable to indicate whether the respective test was published in a journal that belongs to the 'Top 5'.[24] A total of 31,104 tests belong to this category.

The rates of strong reporting errors over time broken down by the two hypotheses and distinguished by our four data sets are given in Figure 3 and Figure 4. Articles published in top 5 journals tend to be less frequently afflicted by reporting errors from 2012 and beyond, especially for the main tests, while visual inspection of the difference between articles with and without data and code availability policy is ambiguous. In general, reporting errors tend to decrease over time.

[24]These are JPE, RESTUD, QJE, AER and Econometrica. The latter two do not offer their articles in HTML and, therefore, are not part of our analysis.



(a) All tests

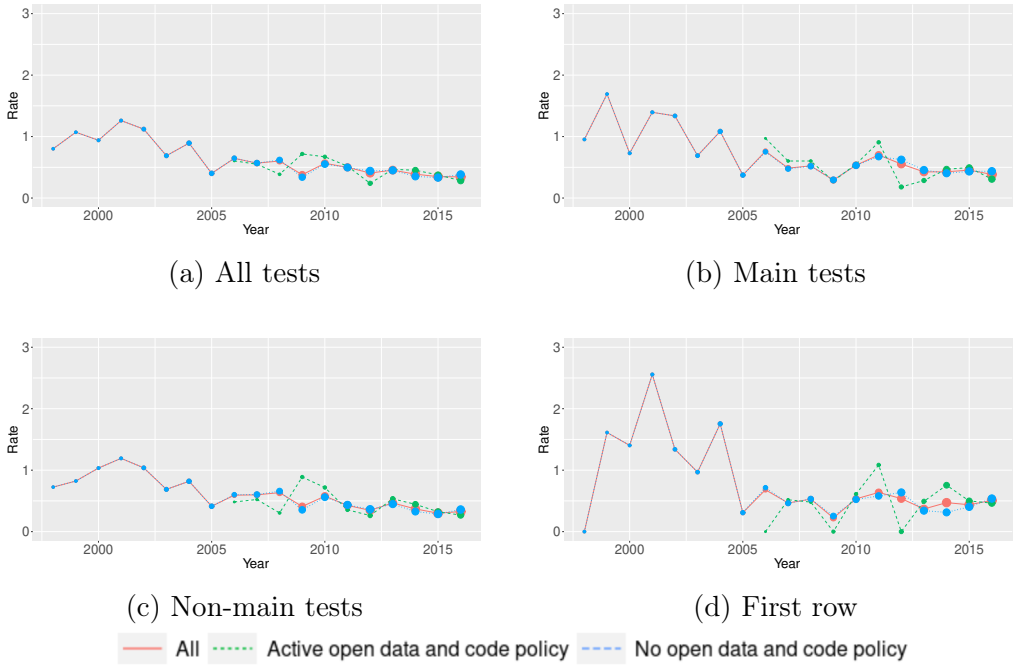
(b) Main tests

(c) Non-main tests

(d) First row

Notes: Rate of strong reporting errors among all data sets distinguished by top 5. Main tests refers to a subset of tests that appear in the first three rows of tables that are neither robustness checks nor appear in the appendix. First row refers to a subset of tests that appear in the first row of tables that are neither robustness checks nor appear in the appendix. The larger a dot, the more tests are in the respective sample. Sizes are not comparable across subfigures.

Figure 3: Rates of strong reporting errors for all data sets distinguished by top 5



(a) All tests

(b) Main tests

(c) Non-main tests

(d) First row

Notes: Rate of strong reporting errors among all data sets distinguished by open data and code policy. Main tests refers to a subset of tests that appear in the first three rows of tables that are neither robustness checks nor appear in the appendix. First row refers to a subset of tests that appear in the first row of tables that are neither robustness checks nor appear in the appendix. The larger a dot is, the more tests are in the respective sample. Sizes are not comparable across subfigures.

Figure 4: Rates of strong reporting errors for all data sets distinguished by presence of a data and code policy

4.2 Pre-registered analysis

Table 4 shows our main results using a logistic regression at the test level. We pick two different dependent variables. First, we choose a dummy variable that indicates if a test is afflicted by a strong overstated reporting error; second, we choose a dummy variable that indicates if a test is afflicted by a strong understated reporting error. Each regression is executed with and without controls for all four of our data sets, resulting in a total of 16 regressions. The second column reflects our pre-registered hypotheses. The other columns function as exploratory analyses. We cluster the standard errors at the article level.

As outlined in the pre-analysis plan, we used a double lasso approach to select control variables (Urminsky et al. 2016). The results are given as odds-ratios where a coefficient that is greater than one indicates a positive influence, and a value that is less than one indicates a negative influence. We report the unadjusted p-values in parentheses and the FDR-adjusted p-values in brackets in order to control for multiple hypotheses testing.[25]

With regards to hypotheses 1a and 1b, the point estimates for the effect of data and code journal policies on the prevalence of reporting errors are all negative except for strong understated reporting errors in the data set containing only the first row in tables that are neither robustness checks nor part of the appendix. This gives indication that such journal policies might reduce the prevalence of strong reporting errors. Looking at the second column, we find that there is a suggestive difference from zero at the 10 % level for strong overstated reporting errors, but not for strong understated reporting errors when considering the unadjusted p-values. The effect size indicates that the chance of a strong overstated reporting error is reduced by 18.1 % when a test is published in a journal with an active open data and code policy. The exploratory regressions in the other columns yield a suggestive difference from zero for the uncontrolled version of the model in the first column and for the uncontrolled versions of the models for non-main tests. We interpret these findings as support for hypothesis 1a while hypothesis 1b is not supported, but it should be emphasized that none

[25] As we have two main variables (*Data and code required* and *Top 5*) and also two dependent variables (strong overstated and strong understated reporting error), we adjust the p-values for four hypotheses.

of the effects is statistically significant once multiple hypothesis testing is taken into account.

With regard to hypotheses 2a and 2b, we find mixed results concerning the direction of the effect. In only 5 out of 16 regressions, publishing a test in the top 5 economic journals could reduce the chance of a strong reporting error. Throughout all models, we find no evidence for statistical significance. These models control for journal rank using the Scimago Journal Rank and this might capture a potential effect of the top 5 journals. When excluding journal rank from the regression suggestive evidence for a positive association between top 5 and strong overstated reporting errors considering all tests is found (cf. Table B.11 in the Online Appendix). This regression is exploratory as we pre-registered journal rank to be a control variable. We conclude that there is no evidence in our data to support hypothesis 2b but some suggestive indication in support of 2a is found.

The coefficients for the control variables can be found in Table B.9 and Table B.10 in the Online Appendix for the strong overstated and the strong understated reporting errors, respectively. An explanation of the control variables can be found in Table B.6 in the Online Appendix. The usage of clustered standard errors reduces the chance of a test being afflicted with a strong overstated reporting error while increasing the chance of a test being afflicted with a strong understated reporting error. Interestingly, the individual chance per test of being afflicted with both types of strong reporting errors is negatively correlated to the number of tests per article. Time appears to have an ambiguous influence on the prevalence of reporting errors. While the effect seems to be negative for overstated reporting errors in non-main tests and for understated reporting errors in the first row of a table, the effect is positive for overstated reporting errors in the first row. Reporting coefficients along with p-, t- or z-values is negatively correlated with overstated reporting errors. The latter also holds for tests in tables with standard significance levels.[26] Results for the number of authors are not statistically significant.

Additionally, we performed the same regressions, but at the article level (cf. Table B.12 in the Online Appendix) as outlined in our pre-analysis

[26]We define standard significance levels as 0.1, 0.05 and 0.01 as around 85 % of all tests in our data measure statistical significance using these levels.

Table 4: Logistic regression at the test level

	All		Main tests		Non-main tests		First row	
Strong overstated								
Data and code required	0.7401	0.8190	0.7815	0.8468	0.7111	0.8106	0.9532	0.9625
	(0.0109)	(0.0786)	(0.1179)	(0.3057)	(0.0214)	(0.1434)	(0.8431)	(0.8674)
	[0.0436]	[0.3144]	[0.3034]	[0.4527]	[0.0856]	[0.5736]	[0.9940]	[0.8674]
Top 5	1.1044	1.2441	0.8460	1.3668	1.2704	1.1785	1.0680	2.1725
	(0.6049)	(0.2914)	(0.5737)	(0.3812)	(0.3116)	(0.5173)	(0.8793)	(0.2039)
	[0.8059]	[0.3885]	[0.7649]	[0.4527]	[0.6232]	[0.6969]	[0.9940]	[0.5778]
Controls	No	Yes	No	Yes	No	Yes	No	Yes
McFadden's Pseudo R^2	0.0009	0.1208	0.0009	0.1229	0.0011	0.1210	0.0000	0.1231
Strong understated								
Data and code required	0.8953	0.8794	0.7842	0.7810	0.9437	0.9200	0.9983	1.1322
	(0.3489)	(0.2195)	(0.1517)	(0.1442)	(0.6874)	(0.5227)	(0.9940)	(0.6211)
	[0.6978]	[0.3885]	[0.3034]	[0.4527]	[0.8050]	[0.6969]	[0.9940]	[0.8281]
Top 5	0.9507	1.1697	0.9917	1.3043	0.9393	1.1041	0.8394	1.7956
	(0.8059)	(0.4480)	(0.9787)	(0.4527)	(0.8050)	(0.7052)	(0.6992)	(0.2889)
	[0.8059]	[0.4480]	[0.9787]	[0.4527]	[0.8050]	[0.7052]	[0.9940]	[0.5778]
Controls	No	Yes	No	Yes	No	Yes	No	Yes
McFadden's Pseudo R^2	0.0001	0.1114	0.0002	0.0970	0.0001	0.1206	0.0006	0.0939
Observations	578,132	578,132	172,040	172,040	406,092	406,092	65,810	65,810

Notes: Logistic regression with double lasso approach for variable selection of controls at the test level. Standard errors are clustered at the article level and based on 5,000 bootstrap replicates. Odds ratios are given with p-values in parentheses and FDR-adjusted p-values in brackets. Intercept not reported. Information on the control variables is given in Table B.6 in the Online Appendix. Main tests refers to a subset of tests that appear in the first three rows of tables that are neither robustness checks nor appear in the appendix. First row refers to a subset of tests that appear in the first row of tables that are neither robustness checks nor appear in the appendix.

plan. In this robustness check, the dependent variable is one if an article contains at least one strong overstated or strong understated reporting error, respectively. We find similar significance levels and point estimates for the uncontrolled models regarding hypothesis 1a. Hypothesis 1b remains unconfirmed. Regarding hypothesis 2b we find a suggestive indication for a negative association for the uncontrolled model for the non-main test.

4.3 Exploratory difference-in-differences

The pre-registered analysis indicates that data and code policies may reduce the risk of strong reporting errors. At the same time, we noticed a negative time trend in the rates of reporting errors while there is an uptake of journals that introduce such policies over time. We perform an exploratory quasi-experimental difference-in-differences design to shed further light on causality. We interpret an open data and code policy as a treatment at the journal level resulting in the following two-way fixed effects model:

$$y_{ijt} = \beta_0 d_{jt} + \beta_1 x_{ijt} + u_i + v_t + \epsilon_{ijt}, \tag{5}$$

where y_{ijt} is the dependent variable that is one if a test i in journal j in time period t is afflicted with a strong overstated or understated reporting error. The binary variable d_{jt} depicts whether a journal j was treated in time period t. x_{ijt} is a set of control variables that include dummy variables for the number of authors, the number of tests per article, dummy variables for the type of test, for the usage of standard significance levels, as well as for the prevalence of clustered standard errors in the corresponding table and the occurrence of another strong reporting error within the same article. The variable selection is mostly based on the significant control variables found in the main regression. u_i and v_t are journal and time fixed effects, respectively.

Figure 5 depicts that there is heterogeneity in the treatment timing. This results in a bias when applying the traditional two-way difference-in-differences design (e.g., Borusyak et al. 2021; Sun and Abraham 2021). Hence, we follow Askarov et al. (2022) who analyze the effect of open data policies on publication bias using the imputation estimator of Borusyak et al. (2021) (BJS).

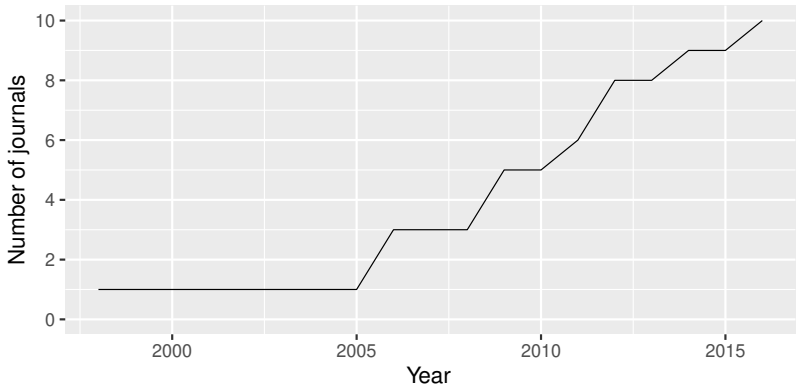


Figure 5: Number of journals with open data and code policies

We consider biannual periods and a minimum of ten articles per journal per year to increase power. This new data set starts in 2003 and comprises the journals *European Economic Review* (EER), *Economic Journal* (EJ), JODE, JOF, *Journal of Financial Economics* (JOFE), *Journal of International Economics* (JOIE), JOLE and JPUE. This results in four treated journals and four non-treated journals while the not yet treated journals

also function as part of the control group. Table B.13 in the Online Appendix gives an explanation of the journal selection.

Table 5 shows the result of our exploratory analysis. We report the coefficient of the policy variable as well as its p-value. Additionally, we report the p-value of the pre-trend test which tests the necessary parallel trends assumption of the imputation estimation of Borusyak et al. (2021). The assumption holds if the pre-trend time coefficients are jointly insignificant. In total, we see that all point estimates are negative. In comparison to Table 4 the effect size is smaller. This might have two reasons. First, while we depict odds-ratios in our main regression, the imputation estimator returns total probabilities.[27] Additionally, by having a specific control group that controls for unspecified differences, we might diminish the effect size. The results are suggestively different from zero at the 10 percent level when considering all tests, and when considering main or non-main tests for the model with journal and year fixed effects as well as with controls. The respective pre-trend tests confirm the findings. The results for the first row remain insignificant. This is in line with the results of Table 4. We also see suggestive indication for a negative effect for the non-main tests in the context of strong understated reporting errors. Overall, we see causal evidence for hypothesis 1a.

As a first robustness check we allow an anticipation effect of one time period. Askarov et al. (2022) show that up to two years prior to the implementation of a data-sharing policy, authors begin to share their data. This suggests that authors change their workflow in anticipation of a data-sharing policy, e.g. when policies are discussed at conferences beforehand, or that at least the implementation in one journal leads to spillover effects to other journals.[28] The results of this robustness check can be observed in the second half of Table 5. The overall picture remains the same for strong overstated reporting errors with p-values that are a little larger on average. This also holds for the negative coefficient for the non-main tests

[27] The imputation estimator of Borusyak et al. (2021) uses the ordinary least squares technique, even though we have a binary dependent variable. This results in a linear probability model. The coefficient is still interpretable, but predictions may be outside the interval [0, 1] (Norton 2007; Askarov et al. 2022).

[28] All journals of the American Economic Association started to implement data-sharing policies in 2005. As these journals do not provide HTML documents, they are not part of our sample.

Table 5: BJS imputation estimator at the test level

	All		Main tests		Non-main tests		First row	
No anticipation								
Strong overstated								
Data and code required	-0.0011	-0.0013	-0.0014	-0.0018	-0.0009	-0.001	-0.0015	-0.0018
p-value	(0.0663)	(0.0191)	(0.1287)	(0.0481)	(0.1708)	(0.0996)	(0.3059)	(0.2191)
Pre-trend test	[0.6242]	[0.9396]	[0.6439]	[0.7143]	[0.6035]	[0.8166]	[0.7362]	[0.6912]
Strong understated								
Data and code required	-0.0007	-0.0008	0.0007	0.0005	-0.0014	-0.0013	0.0008	0.0007
p-value	(0.2134)	(0.1190)	(0.3283)	(0.5122)	(0.0578)	(0.0370)	(0.5508)	(0.6152)
Pre-trend test	[0.0563]	[0.3088]	[0.1341]	[0.2002]	[0.1541]	[0.6674]	[0.0429]	[0.0320]
One time period anticipation (robustness check)								
Strong overstated								
Data and code required	-0.0013	-0.0012	-0.0020	-0.0021	-0.0011	-0.0008	-0.0017	-0.0018
p-value	(0.0671)	(0.0684)	(0.0800)	(0.0556)	(0.1855)	(0.2674)	(0.2961)	(0.2807)
Pre-trend test	[0.2860]	[0.6120]	[0.9374]	[0.9662]	[0.1799]	[0.2876]	[0.6407]	[0.4661]
Strong understated								
Data and code required	-0.0013	-0.0008	0.0006	0.0006	-0.0020	-0.0014	0.0006	0.0007
p-value	(0.0798)	(0.1902)	(0.5640)	(0.5170)	(0.0213)	(0.0714)	(0.7494)	(0.6984)
Pre-trend test	[0.0716]	[0.3458]	[0.0064]	[0.0970]	[0.2630]	[0.7055]	[0.1223]	[0.1355]
Journal and year effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	340,035	340,035	101,764	101,764	238,271	238,271	39,109	39,109

Notes: BJS imputation estimator at the test level showing the average treatment effect on the treated as probabilities. The dependent variable in the first and third quarter of the table is a dummy that is one if a test is afflicted with a strong overstated reporting error. The dependent variable in the second and fourth quarter of the table is a dummy that is one if a test is afflicted with a strong understated reporting error. Coefficients are probabilities. Standard errors are clustered at the article level. p-values of the coefficients in parentheses and p-values of pre-trend tests with three periods in brackets depicted. Controls include dummy variables for the number of authors, the number of tests per article, dummies for the test type, for the usage of standard significance levels, as well as for the prevalence of clustered standard errors in the corresponding table and the occurrence of another strong reporting error within the same article. The data set comprises the journals EER, EJ, JODE, JOF, JOFE, JOIE, JOLE and JPUE from 2003 to 2016 and defines every two years as one time period, i.e. one year. Main tests refers to a subset of tests that appear in the first three rows of tables that are neither robustness checks nor appear in the appendix. First row refers to a subset of tests that appear in the first row of tables that are neither robustness checks nor appear in the appendix.

in the context of strong understated reporting errors. Hence, our findings are robust with regard to anticipation effects. As a second robustness check, we calculated falsification tests in the same manner as Askarov et al. (2022). These falsification tests shall prove that our findings are not based on chance. The first falsification test is depicted in Table B.14. We rerun our analysis but removed all treated observations and moved the treatment to two time-periods earlier, i.e. four years. In the end, no result was statistically significant, supporting the robustness of our findings. For the second falsification test we permute the treatment variable 1,000 times. Figure C.11 and Figure C.12 show the resulting density plots that confirm that our findings regarding hypothesis 1a continue to be robust.

5 Discussion

When considering all tests, the rate of strong statistical reporting errors is comparably large with 30.89 % at the article level. However, previous literature focuses on the main tests of the respective articles and we find

the rate of strong reporting errors with 14.88 % to be in the same ballpark as previous studies. At the test level, the rate of strong reporting errors is small with 0.51 % and at the lower end of what was found for other fields. Economics articles document substantially more tests per article compared to articles in psychology.[29] Therefore, the low error rate at the test level translates into a larger error rate at the article level due to more possibilities of having a reporting error within a respective article. This suggests that reporting errors are at least not more common in economics than in other disciplines. The study that is most comparable to our study is Pütz and Bruns (2021) and uses a manually coded sample. They find with 21.6 % at the article level and 0.5 % at the test level very similar rates to what we found with DORIS on a large scale. Bruns et al. (2019) find larger error rates with 25 % at the article level and 1.4 % at the test level for *Research Policy* which publishes economics articles but also to a large extent articles from management and other disciplines.

The rate of strong reporting errors is at first sight small with only 0.51 % for the main tests. However, reporting errors may mislead readers. For example, many regression models are usually presented in one table to convince the reader about the robustness of a certain finding. Reporting errors may convey the impression that a specific variable has a statistically significant effect across all models. Specifically, 39.74 % of the tables that are neither robustness checks nor part of the appendix with at least one reporting error in the first three rows have at least 10 % of the main tests being afflicted with a strong reporting error. This corresponds to 2.04 % of all tables. Such reporting errors may pose a challenge to cumulative knowledge production and a threat to evidence-based policy making (cf. Section A.7 in the Online Appendix). In the end, reporting errors are avoidable mistakes. Therefore, we can support Askarov et al. (2022) and Pütz and Bruns (2021) who conclude that reducing reporting errors alone is an important contribution.

We find an excess of overstated reporting errors when focusing on the main tests. This finding is in line with previous literature (e.g., Nuijten et

[29]Based on Nuijten et al. (2016) psychology publishes 15.46 main tests in articles with at least one null-hypothesis significance testing result. We estimate this amount at least 46.79 based on Table 3.

al. 2016; Pütz and Bruns 2021). There might be several reasons for a bias in favor of overstated reporting errors as Nuijten et al. (2016) and Pütz and Bruns (2021) pointed out. First, authors' prior beliefs are usually in favor of the alternative hypothesis, and behavioral biases make it easier to accept a statistically significant finding rather than a non-significant one (Kunda 1990; Bastardi et al. 2011). Second, the overall publication process is biased in favor of statistically significant results (e.g., Ioannidis 2005) with some exceptions (Ioannidis 2023). Third, QRP can be another potential source, as, e.g., John et al. (2011) found in psychology that authors frequently round down p-values in order to let them appear statistically significant. In economics, a pendant would be to add eye-catchers to the results to let them appear statistically significant. As the rate of understated reporting errors is relatively large as well, many errors are likely to be honest mistakes that could be easily found during the review process. An alternative to tackle honest mistakes is to prohibit the use of eye-catchers as was done by the AER. While such a policy tackles reporting errors as a byproduct, it mainly addresses the ubiquitous dichotomization of statistical findings using arbitrary thresholds (Cumming 2014).

Our regression results give a slight indication that the top 5 economics journals might have a higher probability of strong overstated reporting errors. This result is surprising as the review process tends to be longer (Ellison 2002) and, therefore, can be assumed to be more rigorous. Additionally, authors may more carefully check manuscripts sent to top journals. In contrast, the reward of publishing in the top 5 is high (Heckman and Moktan 2020; Ductor et al. 2020) which may increase the prevalence of reporting errors. Apparently, the current review process is not sufficient to clear out these mistakes. One way to improve the review process for all journals would be the automatic check for reporting errors, using tools like DORIS or *statcheck*.[30] Note, that Brodeur et al. (2020) find that tests in the top 5 are not more or less likely to be statistically significant.

The probability of a strong overstated reporting error is reduced for journals with open data and code policies, although evidence for under-

[30]In psychology, *statcheck* is already in use at *Psychological Science* and the *Journal of Experimental Social Psychology* to assist in the peer review process (Nuijten et al. 2017).

stated reporting errors is ambiguous. Data and code availability policies are a key instrument to foster replicability (Munafò et al. 2017) and research credibility (Askarov et al. 2022). These policies tackle reporting errors as a byproduct, most likely by incentivizing authors to double check their results for consistency with the replication package before publication. Our finding that these policies are more prone to reduce overstated than understated reporting errors could be due to significant findings being more in line with authors' expectations and hence are not double checked (Nuijten et al. 2016; Pütz and Bruns 2021). Mandatory data and code policies are also heterogeneous[31] and several studies suggest that even though such policies are in place, some authors do not publish neither data nor code (e.g., Müller-Langer et al. 2017; Vlaeminck and Herrmann 2015; Christensen and Miguel 2018; Askarov et al. 2022; Christensen et al. 2019).[32] In the end, open data and code policies are necessary but not sufficient to avoid mistakes and QRP. We suggest enforcing these policies when enacted and also verifying the published data and code before publication. The effort to publish data and code is beneficial to authors, as publicly available data sets may increase citations (Christensen et al. 2019). Nonetheless, we advise to secure confidentiality and privacy. This seems trivial, but Christensen and Miguel (2018) report several cases in which researchers violated the right of privacy by publishing their data.

We might underestimate the effect size of data and code policies as authors might have published their data and code and submitted to journals with open data and code policies in the first place but were finally published in journals without these policies. We have no reason to believe in an overestimation of the effect, as this would mean that authors consciously submit studies containing reporting errors to journals without open data and code policies.

The overall error rate seems to decline slowly over time as can be seen in Figure 3 and Figure 4 and in the regressions, at least for overstated

[31] Vlaeminck and Herrmann (2015) shows that some policies require authors to hand in data before the first submission, while some policies give more leeway. We do not distinguish between these policies.

[32] Reasons for exemptions could be the usage of confidential data but also the non-willingness to share data in order to avoid that someone else uses the data to publish new findings first.

reporting errors for the non-main tests and understated for the first row in a table, which is in line with Nuijten et al. (2016) who find that error rates in psychology have been stable or even declining over time.

In general, DORIS is a powerful tool that automates the data collection and flagging process despite the heterogeneous publication style in economics. It automatically detects different significance levels, collects meta-data about the article and the respective table, and also removes tests where the null hypothesis is not a null effect. DORIS facilitates meta-research in economics[33] and can be used in the review process.

Although we collected a relatively large sample, our study has limitations. DORIS is technically limited and does not recognize every statistical test in every article, e.g., DORIS does not collect tests in tables with fewer than three tests or if significance levels are only marked by bold characters (cf. Section A.4 of the Online Appendix for a comprehensive list). Nevertheless, like written in Section 2.2.2 comparing our data with the data of Brodeur et al. (2016) yields that we find 28 % to 70 % more tests per article, which is plausible, as we consider every test and not only main tests. For comparison, *statcheck* detects 67.5 % of all tests (Nuijten et al. 2016). Second, the FDR of DORIS for detecting tests is 1.2 %. While this FDR requires manually checking flagged tests for the analysis of reporting errors, other meta-research (e.g., analysis of biases) may consider an FDR of 1.2% to be fairly acceptable. The in-depth work of manually reviewing more than 13,900 tests ensures that we minimized the risk of erroneously accusing authors of reporting errors. We suggest that authors always state if they use standard errors or t- or p-values instead, or if they use one-sided or two-sided tests, especially if they diverge from the norm of reporting two-sided tests with standard errors.

6 Conclusion

We analyze the prevalence of statistical reporting errors in more than 578,000 tests from the top 50 economics journals available in HTML (31 journals) using an automated procedure called DORIS (**D**iagnosis **O**f **R**eporting

[33]DORIS could also be extended to other disciplines.

errors **I**n **S**craped tables). This procedure automatically scrapes the articles from the journal web page, extracts its tables, and interprets the values based on text-mining techniques. We find that 0.46 % of all tests are afflicted with a strong reporting error where either the eye-catcher or the calculated p-values signals statistical significance, but the respective other does not. 30.89 % of all articles have at least one strong reporting error. Focusing on the main tests of the respective articles, these error rates become 0.51 % and 14.88 % at the test and article level, respectively.

We find indications of a systematic bias towards statistically significant reporting errors, reflecting a preference for statistically significant findings in the research and publication process. Reporting errors can mislead readers and 1.37 % of all tables in our sample have more than 10 % of the tests being afflicted with strong reporting errors. This number increases to 2.04 % when considering only main tests.

We find suggestive indication that the top 5 economics journals have a higher probability of strong overstated reporting errors even though they have a longer, and hence presumably more rigorous review process. The review process could be improved for all journals by using automated procedures such as *statcheck* and DORIS as an automated tool to flag potential errors before publication. This tool could be used by reviewers immediately after submission as a first check or by authors themselves before submission.[34]

Moreover, there is evidence that mandatory open data and code policies reduce the probability of a test being afflicted with a strong overstated reporting error causally, but evidence on strong understated reporting errors is ambiguous. This finding places even more emphasis on the need for such policies in addition to their benefits in terms of replicability. Although these policies have markedly increased over time to nearly 20 % of all empirically-oriented economics journals having a mandatory data availability policy (Vlaeminck 2021), we suggest that all journals implement not only data availability policies but also code availability policies.

In summary, the publication process could play a key role in detecting statistical reporting errors. Due to open data and code policies, authors

[34] As a HTML document is only produced at the very end of the publication process, we are currently working to adapt DORIS to be applicable to LaTeX files as well.

might check their results more carefully while giving reviewers and other scientists the possibility to verify their research. Stodden et al. (2016) and Artner et al. (2020) suggest a citation system that lists data sets used, code packages, software, and algorithms used in the reference section. This could increase incentives to share data and code. As research becomes more open, the prevalence of reporting errors is likely to diminish.

Another promising and sustainable solution to reduce the prevalence of reporting errors may be to invest more effort in statistical education. The more statistical background the researchers and reviewers have, the fewer errors may be produced. Learning statistical tools that facilitate or automate the transition from statistics to typesetting software (e.g., R packages stargazer and texreg or R Markdown in general) reduces the most common reason for reporting errors based on the survey carried out by Pütz and Bruns (2021).

Future research could address how the actual availability of data and code is associated with the presence of statistical reporting errors. DORIS could be extended to also cover confidence intervals that are increasingly used in economics, and the data compiled by DORIS is likely to be valuable for future meta-research in economics.

Bibliography

Artner, Richard et al. (2020). "The reproducibility of statistical results in psychological research: An investigation using unpublished raw data." In: *Psychological Methods*. ISSN: 1082-989X. DOI: 10.1037/met0000365.

Askarov, Zohid et al. (2022). "The Significance of Data-Sharing Policy". In: *Journal of the European Economic Association*. ISSN: 1542-4766. DOI: 10.1093/jeea/jvac053.

Bach, Laurent et al. (2023). "Dividend Taxes and the Allocation of Capital: Comment". In: *American Economic Review* 113.7, pp. 2048–2052. ISSN: 0002-8282. DOI: 10.1257/aer.20221432.

Bakker, Marjan and Jelte M. Wicherts (2011). "The (mis)reporting of statistical results in psychology journals." In: *Behavior research methods* 43.3, pp. 666–78. DOI: 10.3758/s13428-011-0089-5.

Bakker, Marjan and Jelte M. Wicherts (2014). "Outlier Removal and the Relation with Reporting Errors and Quality of Psychological Research". In: *PLoS ONE* 9.7, e103360. DOI: 10.1371/journal.pone.0103360.

Bastardi, Anthony, Eric Luis Uhlmann, and Lee Ross (2011). "Wishful Thinking". In: *Psychological Science* 22.6, pp. 731–732. ISSN: 0956-7976. DOI: 10.1177/0956797611406447.

Berle, David and Vladan Starcevic (2007). "Inconsistencies between reported test statistics and p-values in two psychiatry journals". In: *International Journal of Methods in Psychiatric Research* 16.4, pp. 202–207. ISSN: 1557-0657. DOI: 10.1002/mpr.225.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess (2021). "Revisiting event study designs: robust and efficient estimation". In: *arXiv*. DOI: 10.48004/wp.cem.2022.1122. eprint: 2108.12419.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes (2020). "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics". In: *American Economic Review* 110.11, pp. 3634–3660. ISSN: 0002-8282. DOI: 10.1257/aer.20190687.

Brodeur, Abel et al. (2016). "Star Wars: The Empirics Strike Back". In: *American Economic Journal: Applied Economics* 8, pp. 1–32. ISSN: 1945-7782. DOI: 10.1257/app.20150044.

Bruns, Stephan B. et al. (2019). "Reporting errors and biases in published empirical findings: Evidence from innovation research". In: *Research Policy* 48, p. 103796. ISSN: 0048-7333. DOI: 10.1016/j.respol.2019.05.005.

Caperos, José Manuel and Antonio Pardo (2013). "Consistency errors in p-values reported in Spanish psychology journals." In: *Psicothema* 25.3, pp. 408–14. DOI: 10.7334/psicothema2012.207.

Card, David and Stefano DellaVigna (2013). "Nine Facts about Top Journals in Economics". In: *Journal of Economic Literature* 51.1, pp. 144–161. ISSN: 0022-0515. DOI: 10.1257/jel.51.1.144.

Christensen, Garret and Edward Miguel (2018). "Transparency, Reproducibility, and the Credibility of Economics Research". In: *Journal of Economic Literature* 56.3, pp. 920–980. ISSN: 0022-0515. DOI: 10.1257/jel.20171350.

Christensen, Garret et al. (2019). "A study of the impact of data sharing on article citations using journal policies as a natural experiment". In: *PLoS ONE* 14.12, e0225883. DOI: 10.1371/journal.pone.0225883.

Clarivate Analytics (2020). *Journal impact factor. Journal Citation Reports 2019*.

Colombo, Matteo et al. (2018). "Statistical reporting inconsistencies in experimental philosophy". In: *PLOS ONE* 13.4, e0194360. DOI: 10.1371/journal.pone.0194360.

Combes, Pierre-Philipp and Laurent Linnemer (2010). "Inferring Missing Citations: A Quantitative MultiCriteria Ranking of all Journals in Economics". In: *HAL* ffhalshs-00520325f.

Cumming, Geoff (2014). "The New Statistics: Why and How". In: *Psychological Science* 25.1. PMID: 24220629, pp. 7–29. DOI: 10.1177/0956797613504966. eprint: https://doi.org/10.1177/0956797613504966. URL: https://doi.org/10.1177/0956797613504966.

Ductor, Lorenzo et al. (2020). "On the Influence of Top Journals". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3580395.

Ellison, Glenn (2002). "The Slowdown of the Economics Publishing Process". In: *Journal of Political Economy* 110.5, pp. 947–993. ISSN: 0022-3808. DOI: 10.1086/341868.

Epskamp, Sasha and Michèle B. Nuijten (2018). *statcheck: Extract statistics from articles and recompute p-values (1.3.1) [R package]*. URL: https://cran.r-project.org/web/packages/statcheck/index.html.

Ercan, Ilker et al. (2017). "Examination of published articles with respect to statistical errors in veterinary sciences". In: *Acta Veterinaria* 67.1, pp. 33–42. DOI: 10.1515/acve-2017-0004.

García-Berthou, Emili and Carles Alcaraz (2004). "Incongruence between test statistics and P values in medical papers". In: *BMC Medical Research Methodology* 4.1, p. 13. DOI: 10.1186/1471-2288-4-13.

Gorajek, Adam and Benjamin Malin (2021). "Comment on "Star Wars: The Empirics Strike Back"". In: *Federal Reserve Bank of Minneapolis* Staff Report No. 629. DOI: 10.21034/sr.629.

Heckman, James J and Sidharth Moktan (2020). "Publishing and Promotion in Economics: The Tyranny of the Top Five". In: *Journal of Eco-*

nomic Literature 58.2, pp. 419–470. ISSN: 0022-0515. DOI: 10.1257/jel.20191574.

Holmes, Tyson H. (2004). "Ten categories of statistical errors: a guide for research in endocrinology and metabolism". In: *American Journal of Physiology-Endocrinology and Metabolism* 286.4, E495–E501. ISSN: 0193-1849. DOI: 10.1152/ajpendo.00484.2003.

Ioannidis, John P. A. (2005). "Why Most Published Research Findings Are False". In: *PLoS Medicine* 2, e124. ISSN: 1549-1277. DOI: 10.1371/journal.pmed.0020124.

— (2023). "Inverse publication reporting bias favouring null, negative results". In: *BMJ Evidence-Based Medicine*, bmjebm–2023-112292. ISSN: 2515-446X. DOI: 10.1136/bmjebm-2023-112292.

John, Leslie K., George Loewenstein, and Drazen Prelec (2011). "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling". In: *Psychological Science* 23.5, pp. 524–532. ISSN: 0956-7976. DOI: 10.1177/0956797611430953.

Karadeniz, Pinar Gunel et al. (2019). "Statistical errors in articles published in radiology journals". In: *Diagnostic and Interventional Radiology* 25.2, pp. 102–108. ISSN: 1305-3612. DOI: 10.5152/dir.2018.18148.

Kunda, Ziva (1990). "The Case for Motivated Reasoning". In: *Psychological Bulletin* 108.3, pp. 480–498. ISSN: 0033-2909. DOI: 10.1037/0033-2909.108.3.480.

Matray, Adrien and Charles Boissel (2023). "Retraction of "Dividend Taxes and the Allocation of Capital"". In: *American Economic Review* 113.7, pp. 2053–2054. ISSN: 0002-8282. DOI: 10.1257/aer.113.7.2053.

Müller-Langer, Frank, Benedikt Fecher, and Dietmar Harhoff (2017). "The Economics of Replication". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.2914225.

Munafò, Marcus R. et al. (2017). "A manifesto for reproducible science". In: *Nature Human Behaviour* 1.1, p. 0021. ISSN: 2397-3374. DOI: 10.1038/s41562-016-0021. URL: https://doi.org/10.1038/s41562-016-0021.

Norton, Edward C. (2007). *Lecture notes in Interaction Terms in Logit and Probit models*.

Nosek, Brian A. et al. (2021). "Replicability, Robustness, and Reproducibility in Psychological Science". In: *Annual Review of Psychology* 73.1,

pp. 1–30. ISSN: 0066-4308. DOI: 10.1146/annurev-psych-020821-114157
.

Nuijten, Michèle B. et al. (2016). "The prevalence of statistical reporting errors in psychology (1985–2013)". In: *Behavior Research Methods* 48, pp. 1205–1226. ISSN: 1554-351x. DOI: 10.3758/s13428-015-0664-2.

Nuijten, Michèle B. et al. (2017). "The Validity of the Tool "statcheck" in Discovering Statistical Reporting Inconsistencies". DOI: 10.31234/osf.io/tcxaj.

Olken, Benjamin A. (2015). "Promises and Perils of Pre-analysis Plans". In: *Journal of Economic Perspectives* 29.3, pp. 61–80. DOI: 10.1257/jep.29.3.61. URL: https://www.aeaweb.org/articles?id=10.1257/jep.29.3.61.

Pütz, Peter and Stephan B. Bruns (2021). "The (non-)significance of reporting errors in econmics: Evidence from three top journals". In: *Journal of Economic Surveys* 35.1, pp. 348–373. ISSN: 0950-0804. DOI: 10.1111/joes.12397.

Stodden, Victoria et al. (2016). "Enhancing reproducibility for computational methods." In: *Science (New York, N.Y.)* 354.6317, pp. 1240–1241. ISSN: 0036-8075. DOI: 10.1126/science.aah6168.

Sun, Liyang and Sarah Abraham (2021). "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects". In: *Journal of Econometrics* 225.2, pp. 175–199. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2020.09.006.

Urminsky, Oleg, Christian Hansen, and Victor Chernozhukov (2016). "Using Double-Lasso Regression for Principled Variable Selection". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.2733374.

Veldkamp, Coosje L. S. et al. (2014). "Statistical Reporting Errors and Collaboration on Statistical Analyses in Psychological Science". In: *PLoS ONE* 9.12, e114876. DOI: 10.1371/journal.pone.0114876.

Vlaeminck, Sven (2021). "Dawning of a new age? Economics journals' data policies on the test bench". In: *LIBER Quarterly: The Journal of the Association of European Research Libraries* 31.1, pp. 1–29. ISSN: 2213-056X. DOI: 10.53377/lq.10940.

Vlaeminck, Sven and Lisa-Kristin Herrmann (2015). "Data Policies and Data Archives: A New Paradigm for Academic Publishing in Economic Sciences?" eng. In: *New Avenues for Electronic Publishing in the Age of*

Infinite Collections and Citizen Science. Proceedings of the 19th International Conference on Electronic Publishing. Ed. by Milena Schmidt Birgit Dobreva. Amsterdam; Amsterdam: IOS Press, pp. 145–155. DOI: 10.3233/978-1-61499-562-3-145. URL: http://hdl.handle.net/10419/121278.

Weinerová, Josefína, Dénes Szűcs, and John P. A. Ioannidis (2022). "Published correlational effect sizes in social and developmental psychology". In: *Royal Society Open Science* 9.12, p. 220311. ISSN: 2054-5703. DOI: 10.1098/rsos.220311.

Wicherts, Jelte M., Marjan Bakker, and Dylan Molenaar (2011). "Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results". In: *PLoS ONE* 6.11, e26828. DOI: 10.1371/journal.pone.0026828.

Statistical reporting errors in economics

Online Appendix

Stephan B. Bruns[1,2,3], Helmut Herwartz[1], John P. A. Ioannidis[3], Chris-Gabriel Islam[1,2,4,†] and Fabian H. C. Raters[1]

[1]Georg August University of Göttingen, Germany
[2]Hasselt University, Belgium
[3]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, USA
[4]Federal Statistical Office of Germany
[†]Corresponding author: chris-gabriel.islam@uni-goettingen.de

Table of Contents

List of Tables

List of Figures

A DORIS

A.1 Summary of DORIS

The scraping part of DORIS is implemented in R (Version 3.6.3) and results in HTML files of the articles, CSV files of the tables, and text files of the respective table notes. These data are the basis for the interpreter part of DORIS that is implemented in Python (Version 3.8.2).

The rule-based interpreter is based on the following main principles: First, tables with eye-catchers are identified mainly by comparing cell contents with a list of possible eye-catchers. This list includes the following signs: *+†‡abc §$^#. Second, different table styles are identified, such as standard errors below the coefficients or standard errors next to the coefficients. In total, we distinguish between 14 different table styles. Third, table notes are used to link eye-catchers to significance levels using a huge set of regular expressions. Fourth, we gather the meta-data for each test, e.g., the usage of clustered standard errors in the table or the usage of non-linear models in the article using the regular expressions given in Table B.7 in the Online Appendix. Where possible, we tried to exclude footnotes and references from the text-mining procedure. Some meta-data were merged using data from the Web of Science, e.g. the number of references or a dummy for open access articles. Finally, we perform plausibility checks, e.g., we check if DORIS finds p-values that are not between 0 and 1 or if the position of some tests in a single table intersects. If a test in a table is marked as not plausible, we remove the entire table from the sample. For the final extraction of the data, we relied on multiprocessing on a server with 128 cores, which resulted in an extraction duration of around 24 hours.

Note that we manually removed three articles from the sample as DORIS misinterpreted some very complicated tables in it. Additionally, we manually removed 44 articles because we found that they were not real HTML files, but HTML files with an embedded PDF file or only a PDF file available for download.

A.2 Flagging based on degrees of freedom

First, in Figure C.1 we depict an example of an overstated reporting error with a test labeled incorrectly significant at the 0.1 level. The red interval represents the range of p-values that are consistent with the reported eye-catcher (0.1 level but not 0.05 level of significance). The possible significance values based on the redundant information (e.g., coefficients with standard errors) are represented by the green interval. Note that the reported statistical information results in an interval due to rounding uncertainty. The position of the green interval further depends on the degrees of freedom. It moves from right to left as the degrees of freedom increase.

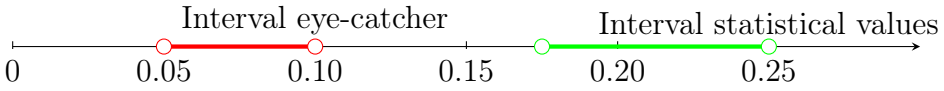


Figure C.1: Overstated reporting error: Intervals of p-values

We search for the maximum number of degrees of freedom (including infinity) df_{crit}^{over} that still ensures that the green and red intervals are disjoint. A strong overstated reporting error occurs when df_{crit}^{over} is defined and when the estimated degrees of freedom are smaller than df_{crit}^{over}.

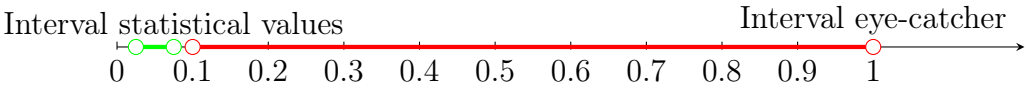


Figure C.2: Understated reporting error: Intervals of p-values

Next, in Figure C.2 we depict an example of an understated reporting error with a test labeled as non-significant, i.e. the significance value is incorrectly labeled between 0.1. and 1. The green interval moves from left to right as the degrees of freedom decrease. We look for the minimum number of degrees of freedom df_{crit}^{under} that just ensures that the green and red intervals are disjoint. A strong understated reporting error occurs when df_{crit}^{under} is defined and when the estimated degrees of freedom are greater than df_{crit}^{under}.

A.3 Manual control

DORIS has a very small FDR for recognizing tests of 1.2 %. Errors were mostly driven by assuming that standard errors are given in parentheses if neither t- or z-values are mentioned in the table notes. Additionally, some tables that reported summary statistics or a t-test for differences of means were erroneously interpreted by DORIS as regression tables. 15 tests (or around 40 %) of the pairs of numbers that were mistakenly interpreted as a test in the evaluation phase of DORIS described in Section 2.1 were diagnosed as a reporting error. Six out of this 15 wrongly flagged tests were marked as overstated reporting errors, while nine were marked as understated reporting errors.

While the FDR is very small, the rate of reporting errors is too. This is why even a very small FDR constitutes a relevant concern for diagnosing reporting errors. Therefore, we use DORIS as a flagging tool that needs manual supervision rather than a fully automated tool for diagnosing reporting errors.

The following steps describe our procedure for manually controlling the output of DORIS. The control was carried out by one co-author and a research assistant who double-checked each other for difficult cases and for some random observations. Another co-author was consulted for complicated cases. The order of the manual control was randomized at the article level. For every strong reporting error, we checked the following variables (additional information that was not collected by DORIS was also gathered to control for potential flagging mistakes as is explained below):

- core data
  - test-type
  - statistical values
  - eye-catcher
  - significance levels given in the table notes
  - over- or understated reporting error?
- meta-data
  - mentions of first-stage regression in table

  - mentions of one-sided tests in article
  - mentions of clustered standard errors in table notes
  - mentions of certain non-linear models in article
  - mentions of usage of significance levels of the underlying model rather than the presented marginal effects in case of certain non-linear models (not collected by DORIS):
    Set it to TRUE, if the table notes mention such a usage, set it to FALSE if it is not mentioned or if it is only mentioned in the text.
  - mentions of multiple hypotheses-testing in article
  - mentions of robustness checks in table
  - mentions of fixed-effects in article
  - number of pages of article
  - number of authors
  - year of publication

- flagging
  - highest number of observations in table or highest integer
  - highest number of observations in text
  - exact number of observations (not collected by DORIS):
    Check if the observations are given in the table, either numerically or as a formula (e.g., quarterly data from 1991Q1 to 1995Q4). Leave it empty if there are no numbers of observations given or if it is only stated in the text.
  - maximum number of rows or columns with parameters
  - exact number of parameters (not collected by DORIS):
    Count the reported parameters per regression. If the table mentions controls or fixed-effects, collect the exact amount by looking into the text, primarily in the data section, and mark the collected number with a comment in the document.
  - number of clusters (not collected by DORIS):
    If the table notes state the usage of clustered standard errors, collect the exact number of clusters by looking first at the table and then at the text, primarily in the data section, and mark the collected number with a comment in the document.

For the regular expressions used to obtain the meta-data, see Table B.7 in the Online Appendix. A mistake in the core data lead to the removal

of the whole article from the data set. A mistake in the meta-data was manually corrected.

Mistakes in variables concerning the flagging were treated three-fold: If no number of observations is given in the table but DORIS mistakenly collected a number, the whole table was removed from the data set, as flagging cannot be performed without an estimation of the degrees of freedom. If DORIS collected a wrong number and the mistake leads to an excession of the threshold of critical degrees of freedom (cf. Section A.2, the entire article was removed to rule out other potential mistakes. If the mistake does not exceed the threshold of degrees of freedom, the mistake was manually corrected in the data set and the article was kept in the sample. Note that for tables with clustered standard errors and understated reporting errors, the number of clusters instead of the number of observations was used to calculate the exact number of degrees of freedom, as Stata sometimes uses the number of clusters as degrees of freedom (Pütz and Bruns 2021). As the number of clusters is usually far lower than the number of observations, this gives the authors the benefit of the doubt as we flag fewer understated reporting errors when assuming lower degrees of freedom.

Additionally, if possible, we determined the reason for a reporting error, e.g., wrong HTML formatting, wrong information about values in parentheses, etc. These data can be obtained from the authors on request.

A.4 Limitations of DORIS

- Small tables are not recognized (minimum three tests).
- DORIS does not interpret tables with bold eye-catchers.
- DORIS cannot recognize image data (e.g., tables or significance levels that are pasted as images).
- In case of multiple standard errors reported with multiple parentheses, one below the other, DORIS only recognizes the first parentheses.
- Sometimes, the significance level is not given per test individually, but rather for a whole table row. In these cases, DORIS only recognizes the last column.
- For Wiley-published journals, DORIS is not able to distinguish the usual text from footnotes, as they are spread in the text and not bundled at the end of the document.
- DORIS has problems extracting the number of observations if they are given in a complicated format, e.g., "255×4 observations".

A.5 Removing outliers

Table B.1: Derivation of outliers

Quantile	0	0.5	1	1.5	2	2.5	3	4	5
Threshold total errors (above)	103.00	36.76	20.76	14.00	12.00	10.00	8.00	6.00	5.00
...affected articles	0	21	41	59	72	93	117	154	192
Threshold share errors (above) [%]	85.71	25.71	18.92	14.29	11.63	8.97	7.86	6.18	4.84
...affected articles	0	21	41	59	81	101	121	161	202
Total affected articles (above)	0	33	63	88	112	139	173	227	277
Threshold total tests (below)	2	4	6	6	7	8	9	12	13
...affected articles	0	8	31	31	61	77	103	141	178
Threshold share tables with tests (below) [%]	3.12	7.69	8.33	9.09	9.09	10.00	10.00	11.11	11.91
...affected articles	0	12	24	41	41	71	71	109	170
Total affected articles (below)	0	19	54	69	97	140	164	232	308
Remaining errors	5316	3889	3415	3078	2913	2675	2420	2121	1889
Remaining articles	4025	3973	3908	3868	3816	3746	3688	3566	3440
...ratio [%]	100.00	98.71	97.09	96.10	94.81	93.07	91.63	88.60	85.47
Remaining articles that contain at least one error	1296	1263	1233	1208	1184	1157	1123	1069	1019
...ratio [%]	32.20	31.79	31.55	31.23	31.03	30.89	30.45	29.98	29.62

Notes: Only strong reporting errors are considered. Articles that would have been cut from below and that were controlled during the correction of the flagged strong reporting errors were kept in the sample.

Many reporting errors stem from misreporting the numbers in parentheses or in brackets. Holmes (2004) coined this as Reporting Imprecision. This means that the second number of a test, next to the coefficient, is not specified correctly or specified at all, e.g., standard errors are marked as t-values. This reporting error differs substantially from reporting errors that stem from accidentally inserted eye-catchers as the reporting error share in a table increases tremendously when having reporting imprecisions. This high prevalence overshadows the rather low prevalence of other reporting errors in subsequent analyzes. In order to separate this error, which is usually easily distinguished by the reader from real reporting errors, we try to exclude them from our sample. Table B.1 shows how many articles would be removed if we cut the top x percent from our manually corrected sample (cf. Figure C.8 based on either the number of total reporting errors or the share of reporting errors. On the range of all reporting errors, this resembles a cut from above. Furthermore, since many analyses are calculated at the article level or clustered at the article level, we want to exclude articles for which we might assume that DORIS had a low power in detecting statistical tests, i.e. found only a small portion of all tests and almost likely no reporting errors. Therefore, we look at the number of tests found and the share of tables with tests as can be seen in Table B.1. On the range of all reporting errors, this resembles a cut from below.

In order to keep the removal to a minimum while removing most of the

problematic cases mentioned above, we chose the 2.5 % quantile as the threshold for the definition of outliers. Theoretically, this means that we would cut 10 % from our sample. Effectively, we remove only around 8 % of our sample as we keep articles that would have been cut from below but which were already checked in the manual control and as sometimes the same article is removed by different cuts. In total, this means that we remove articles with more than 10 strong reporting errors or 8.97 % share of strong reporting errors and articles with less than 8 tests or less than 10 % of the tables containing tests.

A.6 Definition of main tests

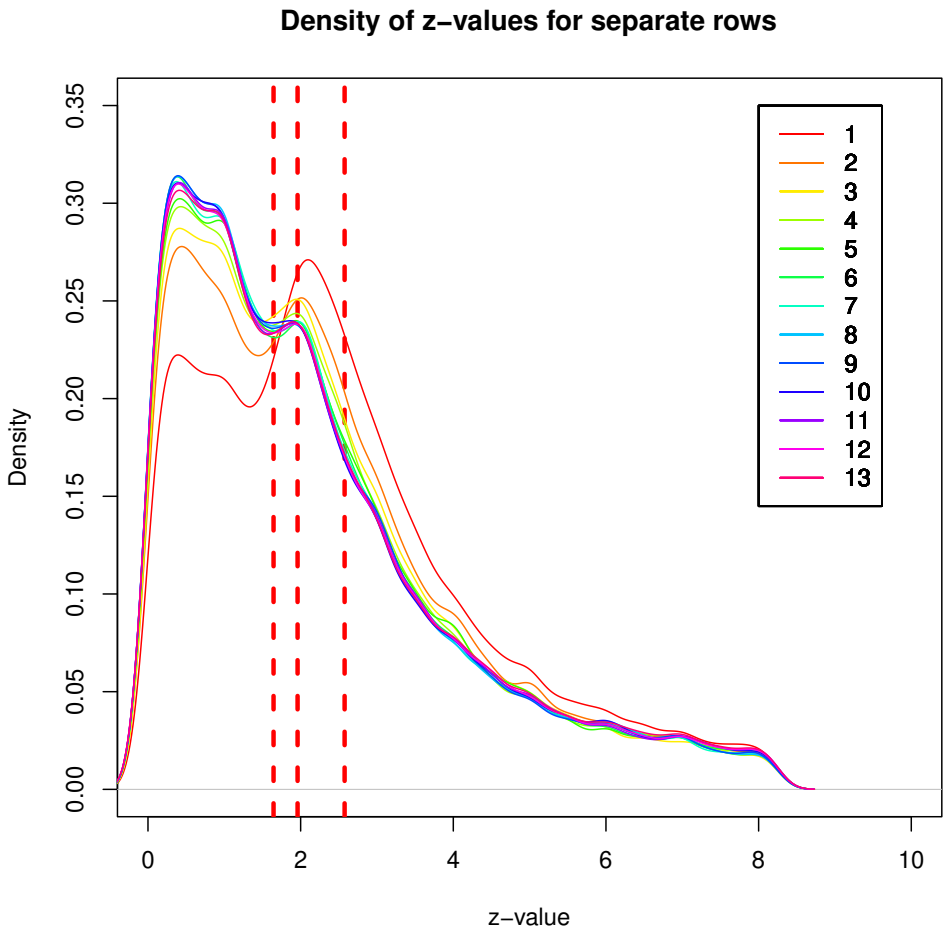


Figure C.3: Density of z-values using different table rows

For the definition of main tests, we assume that in economics regression models are usually reported columnwise and that main tests appears in the top rows of a table. To define the amount of table rows that belong to the main tests on average, we built 13 data sets from our principal data set without outliers (cf. Figure C.8, z-values greater than 10 are cut), each consisting only of tests belonging to the i-th row of tables that are neither robustness checks nor appear in the appendix, where $i \in 1, .., 13$. We chose 13 as the maximum as 80 % of all tables have 13 or fewer rows. Figure C.3 shows the density functions based on z-values for these data sets. The

first red dashed line from the left depicts a z-value of 1.645, i.e. the 10 % significance level, the second depicts 1.96, i.e. the 5 % significance level and the third depicts 2.576, i.e. the 1 % significance level. For the first, second, and third rows we see a clear excess of significant values. The other rows converge to one common line. Gorajek and Malin (2021) find that main tests show, among others, a discontinuity at 1.96, while non-main test do not. Hence, we choose our definition of main tests to include the first three rows of a table to be in line with these findings. As a robustness data set we choose only the first row because we might overestimate the number of main tests using the definition that includes the first three rows.

A.7 Change of statement due to statistical reporting errors

In the following section, we present four examples of articles whose statements might change due to statistical reporting errors. We first quote the text from the respective article, give an explanation about the reporting errors, and the statements that might not hold when considering the reporting errors, and show the respective table afterwards. References are removed for reasons of anonymization.

Example 1

Text

Using a linear probability model (OLS), we regress indicators of future default on ratings and control variables. In Column 1, default in three years is regressed on a dummy for investment-grade rating, Fitch's market share, and an interaction of these. All three variables are highly significant. The coefficient on the investment-grade dummy is negative and significant at the 1% level, implying that firms rated investment grade are less likely to default than those rated noninvestment grade. The interaction is positive, meaning that the difference between investment-grade and speculative-grade default rates falls with competition. The magnitude of this effect is large. [...] In Column 2, we replace the investment-grade dummy with the numerical rating value (using the Hand, Holthausen, and Leftwich, 1992, scale), with similar results. The scale uses finer variation, but at the expense of imposing a particular numerical scale, which could be inappropriate. As it turns out, the fit is slightly better, and the magnitude of the interaction remains large and significant at the 1% level.

Explanation

There is a strong overstated reporting error in the first column, second row, as well as two strong overstated reporting errors in the second column, third, and fifth row. Hence, the statement that "firms rated investment grade are less likely to default than those rated noninvestment grade" might not hold. Additionally, the authors argue that the second column

shows similar results. This might not be true in the face of the reported errors mentioned.

Table

Table 9
Default prediction—the effect of Fitch market share.

Each column presents the coefficient estimates from an ordinary least squares (OLS) regression. Intercepts are not reported. Each observation is one firm-year in which firm-level controls can be identified and the firm is identified as defaulting or not defaulting in three years. The sample period is from 1995 until 2005. Fitch market share is the fraction of bond ratings in an industry-year cell issued by Fitch Ratings. Industries are two-digit level North American Industry Classifications System (NAICS) industries. Firm characteristics are the log of sales, log of book value of assets, cash divided by total assets (and its square), EBITDA (earnings before internet, taxes, depreciation, and amortization)divided by total assets (and its square), cash flow over total assets (and its square), EBITDA over sales (and its square), cash flow over sales (and its square), PPE (property, plant, and equipment) over total assets (and its square), interest expense over EBITDA (and its square), debt over total assets (and its square), and the log of sales and the log of assets, all measured at the end of the previous fiscal year (using accounting data from Compustat). In Column 5, data are averaged by industry-year cell. The standard errors for the coefficient estimates are in parentheses and are clustered by industry × year cell in Column 1 to 4 and heteroskedasticity robust in all columns. Significance at the10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

	Default in three years, OLS (1)	Default in three years, OLS (2)	Default in three years, OLS (3)	Default in three years, OLS (4)	Fraction default in three years by industry-year cell, OLS (5)
IG dummy × Fitch market share	0.089***				
	(0.030)				
IG dummy	−0.033***				
	(0.067)				
Rating × Fitch market share		0.0123***	0.0123***	0.0063*	0.0070***
		(0.070)	(0.001)	(0.0038)	(0.001)
Rating		−0.0045***	−0.0049***		−0.0094**
		(0.0009)	(0.0011)		(0.004)
Fitch market share	−0.080***	−0.253***	−0.229***	−0.116	−0.121***
	(0.028)	(1.250)	(0.085)	(0.080)	(0.023)
Year fixed effects	No	No	Yes	Yes	Yes
Industry fixed effects	No	No	Yes	Yes	Yes
Year and industry FE × rating	No	No	No	Yes	No
Firm controls	No	No	Yes	No	Yes
R^2	0.008	0.001	0.024	0.024	0.571
Number of observations	18,707	18,650	15,661	18,650	189

Example 2

Text

The estimated treatment effect of VC experience, estimated using the Heckman two-step method, is displayed in the second column of Table 8. The significant negative relation between downside protections and VC experience continues to hold.

Explanation

There is a strong overstated reporting error in the first column, first row. Hence, the statement that the "significant negative relation between downside protections and VC experience continues to hold" might not be true.

Table

Table 8
Instrumental variables and Heckman specifications. This table reports instrumental variables and Heckman specifications. Downside protection index (DPI) is the dependent variable (see Table 2 for description). Specification 1 is an IV regression in which VC experience is instrumented with the average experience of all VCs (excluding the actual VC making the investment) in the same state as the company receiving investment. Specification 2 is a Heckman selection model in which the identifying variables (included in the selection equation but not the outcome equation) are interactions of company state and VC state dummies. The selection equation estimates the probability of a match as a function of the controls and identifying dummies; the reported coefficients are estimated treatment (outcome) effects. We restrict the sample in specification 2 to observations in which the VC and company locations are California, Massachussets, Texas, North Carolina or New York. Lambda is Inverse Mill's Ratio. All specifications control for VC and company in same state, serial founder, serial founder with IPO, serial founder with merger, early-stage company, company age, total round amount, number of VCs in round, as well as fixed effects for VC firm state (California, Massachussets, Texas, New York, and other), company state, company industry (Venture Economics 10-level classification), round year, and round number.

Specification	1	2
Identification Method	IV	Heckman–Sorensen
Dependent variable	Downside protection index	Downside protection index
(log) VC experience	−0.363***	−0.286***
	[0.119]	[0.989]
Lambda		−0.687
		[0.525]
Observations	3394	1880
R-squared	0.21	
Sample	Full	California, Massachussetts, Texas, North Carolina, New York
Full set of fixed effects and additional controls	Yes	Yes
Staiger–Stock *F*-statistic	103.99	
Instrument for (log) VC experience	(Log) Average experience of VCs in same state as company (excluding the actual VC)	
Selection equation		Company State * VC State FEs
Number of potential matches		887,300
Diagnostic: *p* (Rho = 0)		0.0668

* Significance at the 10% level.
** Significance at the 5% level.
*** Significance at the 1% level.

Example 3

Text

Contrary to the results of the non-parametric statistical test, the regression picks up a weakly significant difference between participant types, with sellers being about 12 percentage points more likely than buyers to opt for tax evasion, ceteris paribus.

Explanation

There is a strong overstated reporting error in the first column, first row. Hence, the statement that "sellers being about 12 percentage points more likely than buyers to opt for tax evasion" might not be true.

Table

Table 4
Tax decisions, treatment *ENDO*.

Dependent variable	Tax-dishonesty
Seller	0.122*
	(0.075)
p^l	0.001
	(0.001)
p^h	−0.001
	(0.001)
Period	0.002
	(0.003)
Female	−0.214**
	(0.084)
Age	0.010
	(0.011)
Risk factor	−0.021
	(0.035)
Observations	1600

Notes. Probit regressions, marginal effects. Standard errors are in parentheses, clustered by matching group.
** $p < 0.05$.
* $p < 0.1$.

Example 4

Text

These results are reported in Table 6, Panel A.20 In this specification, we find conflicting evidence on the effect of equity capital on credit constraints (columns (1) and (2)) [...] Finally, this time we find that gains (losses) on financial assets are associated with lower (higher) credit constraints (columns (8) and (9)), implying that over the credit cycle, firms' access to credit was higher if they were borrowing from banks whose financial assets were appreciating rather than depreciating in value.

Explanation

There is a strong overstated reporting error in the first row, second column. Hence, the statement of "conflicting evidence on the effect of equity capital on credit constraints" might not be true. Additionally, there are two strong overstated reporting errors in the first row, eighth, and ninth columns. Hence, the statement that "over the credit cycle, firms' access to credit was higher if they were borrowing from banks whose financial assets were appreciating rather than depreciating in value" might not be true.

Table (Panel B is excluded)

Table 6
Transmission of bank balance sheet conditions: 2005 and 2008 samples.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Difference-in-differences 1									
	Finance = equity/assets			Finance = Tier 1 capital ratio			Finance = gains on fin assets		
	Equally-weighted	Branch-weighted	Asset-weighted	Equally-weighted	Branch-weighted	Asset-weighted	Equally-weighted	Branch-weighted	Asset-weighted
Post × Finance	0.075	−0.042	0.011	−0.088	−0.203	−0.070	−0.301	−0.514	−0.562
	(0.033)**	(0.027)*	(0.022)	(0.051)*	(0.043)***	(0.032)**	(0.342)	(0.322)*	(0.354)*
Finance	−0.073	−0.013	−0.033	−0.005	0.044	−0.039	0.316	0.495	0.782
	(0.030)**	(0.028)	(0.026)	(0.028)	(0.028)	(0.024)*	(0.281)	(0.263)*	(0.316)**
Post	−0.354	0.374	0.063	0.864	1.857	0.726	0.180	0.205	0.296
	(0.216)	(0.181)**	(0.144)	(0.444)**	(0.371)***	(0.257)***	(0.102)	(0.095)**	(0.089)***
Inverse Mills' ratio	−0.143	−0.151	−0.151	−0.138	−0.137	−0.133	−0.160	−0.163	−0.150
	(0.056)***	(0.056)***	(0.056)***	(0.055)***	(0.055)***	(0.055)***	(0.058)***	(0.057)***	(0.060)**
Fixed effects	Country Industry								
Observations	5182	5182	5089	5181	5181	5060	5149	5149	5089
Pseudo R-squared	0.08	0.08	0.08	0.08	0.09	0.09	0.08	0.08	0.09

Note: The dependent variable is a dummy variable equal to 1 if the firm is credit constrained. 'Finance' is one of the three balance sheet variables from Table 4. Each finance variable is locality-specific and is constructed by weighting equally (columns labeled "Equally-weighted"), by number of branches (columns labeled "Branch-weighted"), or by subsidiaries' assets (columns labeled "Asset-weighted") the respective financial variable for each parent bank which has at least one branch or subsidiary in that locality. 'Post' is a dummy variable equal to 1 if the observation is in 2008, and to 0 if it is in 2005. 'Non-Affected' is a dummy variable equal to 1 if the respective finance variable declined by less than 1 standard deviation between 2005 and 2008. 'Inverse Mills' ratio' is the inverse of Mills' ratio from the probit model in Table 4 for each respective financial variable. The regressions also include the rest of the independent variables from Table 5 (unreported for brevity). Omitted variables from the probit equation in Table 4 are 'Competition' and 'Subsidized'. The analysis is performed on all firms present either in the 2005 or in the 2008 survey (Panel A), and on all firms present in localities which appeared both in the 2005 and the 2008 survey (Panel B). All regressions include country and industry fixed effects. White (1980) robust standard errors are reported in parentheses, where *** indicates significance at the 1% level, ** at the 5% level, and * at the 10% level. See Appendix 1 for exact definitions. Source: BEEPS (2005 and 2008) and Bankscope (2005 and 2008).

Example 5

Text

We report the results in Table 3. If our previous findings were mainly the result of a simultaneity bias, we should find a significantly positive coefficient on our distance variable in one sample, and possibly an insignificant coefficient in the other. This is not what we find. The distance coefficients remain significant across spread (using the median as cutoff value) and securitization (no, yes) categories. Specifically, the coefficient is 0.1230 ($p < 0.01$) for the low-spread loans (i.e., all-in spread drawn ≤ 4.60) and 0.0922 ($p < 0.05$) for the high-spread loans. (...) Overall, except for this small short-maturity loans category, these results suggest that simultaneity bias is not a primary explanation for our findings.

Explanation

There is a strong overstated reporting error in the first and second columns, first row. Hence, the statement that "simultaneity bias is not a primary explanation for their findings" might not be true.

Table

Table 3
Loan spread, maturity, and securitization.

This table presents estimates from ordered logistic regressions of *Covenant intensity* on *Ln(Distance to lender)*. The control variables are the same as in Model 3 of Table 2. All variables are defined in Table 1. *Loan spread* is the all-in spread drawn (i.e., including fees and the spread that the borrower pays in basis points over the LIBOR for each dollar drawn down under loan commitment). *Loan maturity* is months to maturity. The analysis evaluates syndicated loans. We obtain our information on contract terms, lead arrangers, and participant lenders from DealScan, and our accounting and borrower characteristics from Compustat and CRSP. The sample contains loans with covenant information that were issued by U.S. commercial banks between 2005 and 2008. Standard errors, which are reported in parentheses below the coefficient estimates, are corrected for heteroskedasticity and simultaneous facility-level and borrower-level clustering (Petersen, 2009). We use ***, **, and * to denote that the coefficient estimate is different from zero at the 1%, 5%, and 10% levels (two-tailed), respectively.

	Loan spread (LS)		Loan maturity (LM), in months			Secured loan	
	LS ≤ 4.60	LS > 4.60	LM ≤ 24	24 < LM ≤ 60	LM > 60	No	Yes
Ln(Distance to lender)	0.1230 ***	0.0922 **	0.0904	0.1028 ***	0.1458 **	0.1130 ***	0.0852 ***
	(0.346)	(0.403)	(0.093)	(0.027)	(0.063)	(0.032)	(0.038)
Control variables	Included	Included	Included	Included	Included	Included	Included
Number of observations	4,344	4,064	822	6,260	1,326	4,327	4,081
Number of facilities	1,310	1,803	416	2,175	521	1,346	1,766
Number of companies	803	916	318	1,357	300	813	881
Pseudo *R*-squared	14%	8%	17%	15%	16%	16%	8%

Example 6

Text

When we turn our attention to families at or above the 75th percentile in terms of household size (column 3 of Table 6), we see that health improvements lead to substantial increases in the percent of acreage fallowed by these larger households. While smaller households have a 0.54 percentage point decrease in the percent of land that is fallowed at 100 days on treatment, these larger households have a 0.45 percentage point increase in the percent of land that is fallowed after 100 days of ART. (...) In addition, Column 6 indicates that households with more children under the age of six have a greater increase in their land fallowing as they get healthier, relative to households with less young children (this final result is significant only at the 10% level.)

Explanation

There is a strong overstated reporting error in the third column, first row. Hence, the statement that "smaller households have a 0.54 percentage point decrease in the percent of land that is fallowed at 100 days on treatment" might not be true. Additionally, there is a strong overstated reporting error in the sixth column, first row. Hence, the statement that "households with more children under the age of six have a greater increase in their land fallowing as they get healthier, relative to households with less young children" might not be true.

Table

Table 6
Impact of ART on fallow land: labor endowment effects.

Dependent variable:	(1)	(2)	(3)	(4)	(5)	(6)
	Percent of land that is fallow					
Days on ART	−0.00240	0.00270	−0.00541*	−0.00332	0.00745	0.0112*
	(0.00336)	(0.00823)	(0.00379)	(0.00432)	(0.00690)	(0.00745)
Days on ART * HH size		−0.00433				
		(0.00284)				
Days on ART * (HH size)^2		0.000516**				
		(0.000258)				
Days on ART * HH size>7			0.01000**	0.0121**	0.0116**	0.0145***
			(0.00435)	(0.00499)	(0.00448)	(0.00527)
Days on ART * # children ≤ 5				−0.00499		−0.00653
				(0.00562)		(0.00573)
Days on ART * avg. age					−0.000459*	−0.000497**
					(0.000245)	(0.000248)
Mean (dependent var.):	50.90	50.90	50.90	50.90	50.90	50.90
R-squared	0.063	0.069	0.067	0.068	0.072	0.074
Number of hhn	624	624	624	624	624	624

Standard errors in parentheses. Regressions include household and year fixed effects.
* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

B Tables

Table B.2: Journal overview ordered by IDEAS/RePEc Ranking 12/2018[a]

Journal	Availability of articles in HTML[b]	First volume/issue (year) in data set	Main JEL code[c]
Quarterly Journal of Economics	yes	126/1 (2011)	General
Journal of Economic Literature	no		
American Economic Journal: Macroeconomics	no		
Econometrica	no		
Journal of Political Economy	yes	108/4 (2000)	General
Review of Economic Studies	yes	70/3 (2003)	General
American Economic Journal: Applied Economics	no		
Journal of Finance	yes	53/1 (1998)	Finance
Economic Policy	yes	17/34 (2002)	International
Journal of Economic Perspectives	no		
Journal of the European Economic Association	yes	9/1 (2011)	General
American Economic Review	no		
The Review of Economics and Statistics	no		
Brookings Papers on Economic Activity	no		
Journal of Economic Growth	yes	-[d]	Development/Growth
Journal of Monetary Economics	yes	41/2 (1998)	Macro/Monetary
Journal of Financial Economics	yes	47/2 (1998)	Finance
Annual Review of Economics	yes	1 (2009)	Not included but given "General"
Journal of International Economics	yes	44/1 (1998)	International
American Economic Journal: Economic Policy	no		
IMF Economic Review	yes	43/2 (2002)	International
Journal of Labor Economics	yes	19/2 (2001)	Labour
Journal of Human Resources	no	53/1 (2006)	

Continued on next page

Table B.2 – *Continued from previous page*

Journal	Availability of articles in HTML[b]	First volume/issue (year) in data set	Main JEL code[c]
Journal of Development Economics	yes	55/1 (1998)	Development/Growth
Economic Journal	yes	112/476 (2002)	General
Review of Financial Studies	yes	11/1 (1998)	Finance
Annals of Economics and Finance	no		
Journal of Financial Intermediation	yes	12/1 (2003)	Finance
Review of Economic Dynamics	yes	17/2 (2014)	Macro/Monetary
BIS Quarterly Review	no		Finance
Journal of Applied Econometrics	yes	16/1 (2001)	Econometrics
Journal of Money, Credit and Banking	yes	39/1 (2007)	Macro/Monetary
Journal of Public Economics	yes	67/1 (1998)	Public/Political Science
Journal of Business & Economic Statistics	yes	30/1 (2012)	Econometrics
International Journal of Central Banking	no		
Journal of Urban Economics	yes	66/1 (2009)	Urban/Regional
International Economic Review	yes	43/2 (2002)	General
Experimental Economics	yes	-[d]	Micro/Game Theory
Journal of Economic Surveys	yes	20/4 (2006)	General
RAND Journal of Economics	yes	38/3 (2007)	Industrial Organization
Quantitative Economics	no		
Annual Review of Financial Economics	yes	1 (2009)	Not included but given "Finance"
European Economic Review	yes	42/3-5 (1998)	General
Journal of Environmental Economics and Management	yes	45/1 (2003)	Environmental
World Bank Research Observer	yes	20/2 (2005)	Development/Growth
Journal of Econometrics	yes	86/1 (1998)	Econometrics
Journal of International Money and Finance	yes	17/1 (1998)	International

Continued on next page

Table B.2 – *Continued from previous page*

Journal	Availability of articles in HTML[b]	First volume/issue (year) in data set	Main JEL code[c]
Review of Environmental Economics and Policy	yes	1/1 (2007)	Environmental
Oxford Bulletin of Economics and Statistics	yes	64/1 (2002)	General
Experimental Economics	yes	-[d]	Demography

[a] IDEAS/RePEc Ranking Simple Impact Factors (Last 10 Years) as of December 2018. https://ideas.repec.org/top/old/1812/top.journals.simple10.html, last retrieved on 04.11.2022.
[b] As of January 2019.
[c] Based on Combes and Linnemer (2010). Only for journals with HTML documents.
[d] No scraping due to Springer's conditions.

Table B.3: Previous studies (sample)

Article	Field	Data collection	Strong error rate at article level [%]	Strong error rate at test level [%]	Number of articles	Number of tests
Current study	Economics	DORIS	14.9	0.5	3,677	172,040
Bakker and Wicherts (2014)	Psychology	*statcheck* and manually	15.0	1.1	153	2,667
Berle and Starcevic (2007)	Psychiatry	Manually	9.4	-	96	546
Bruns et al. (2019)	Innovation Economics	Manually	25.0	1.4	101	5,667
Caperos and Pardo (2013)	Psychology	Manually	17.6	2.3	102	1,212
Colombo et al. (2018)	Experimental Philosophy	*statcheck*	6.4	0.5	173	2,573
Ercan et al. (2017)	Veterinary Sciences	Manually	8.8		204	-
García-Berthou and Alcaraz (2004)	Nature and British Medical Journal	Manually	-	0.4	44	244
Karadeniz et al. (2019)	Radiology	Manually	17.8	-	157	-
Nuijten et al. (2016)	Psychology	*statcheck*	12.9	1.4	16,695	258,105
Pütz and Bruns (2021)	Economics	Manually	21.6	0.5	370	30,993
Veldkamp et al. (2014)	Psychology	*statcheck*	20.5	0.8	430	8,105

Source: Own table, partly adapted from Pütz and Bruns (2021).
Notes: For the sake of comparability, the analyzed population is the set of main tests. We and Pütz and Bruns (2021) diagnose strong reporting errors based on the significance levels given in the notes of the respective table (the lowest significance level is mostly $p = 0.1$). All other studies consider a fixed p-value threshold to define strong errors. Bruns et al. (2019) and base their calculations of strong errors on $p = 0.1$ and in the other studies $p = 0.05$ is considered.

Table B.4: Mistakes during the evaluation process of DORIS

Journal	**# mistakes in core data**	**# mistakes in meta-data**	**# tests**
Annual Review of Economics	0	0	100
Annual Review of Financial Economics	0	0	100
European Economic Review	3	5	100
Economic Journal	1	2	100
Economic Policy	3	4	100
International Economic Review	1	8	100
Journal of Business & Economic Statistics	0	0	0
Journal of the European Economic Association	0	3	100
Journal of Money, Credit and Banking	2	7	100
Journal of Monetary Economics	1	2	100
Journal of Applied Econometrics	1	3	100
Journal of Development Economics	2	5	100
Journal of Econometrics	1	3	100
Journal of Environmental Economics and Management	0	6	100
Journal of Economic Surveys	3	2	100
Journal of Finance	4	9	100
Journal of Financial Economics	1	5	100
Journal of Financial Intermediation	6	3	100
Journal of International Economics	2	4	100
Journal of International Money and Finance	0	4	100
Journal of Labor Economics	0	6	100
Journal of Urban Economics	0	6	100
Journal of Political Economy	1	5	100
Journal of Public Economics	1	0	100
Oxford Bulletin of Economics and Statistics	1	12	100
Quarterly Journal of Economics	0	4	100
RAND Journal of Economics	1	11	100
Review of Economic Dynamics	0	0	100
Review of Environmental Economics and Policy	0	0	68
Review of Economic Studies	2	9	100
Review of Financial Studies	1	4	100
World Bank Research Observer	0	13	100
Overall	38	138	3,068

Table B.5: Journal policies and implementation date

Journal	Data required?	Data required since?	Code required?	Code required since?
Annual Review of Economics (ARE)	no		no	
Annual Review of Financial Economics (ARFE)	no		no	
European Economic Review (EER)	yes	2012	yes	2012
Economic Journal (EJ)	yes	2012	yes	2012
Economic Policy (EP)	no		no	
International Economic Review (IER)	yes	2009	yes	2009
Journal of Business & Economic Statistics (JBES)	yes	2011	yes	1993
Journal of the European Economic Association (JEEA)	yes	2011	yes	2011
Journal of Money, Credit and Banking[a] (JMCB)	yes	1996	yes	1996
Journal of Monetary Economics (JME)	no		no	
Journal of Applied Econometrics (JOAE)	yes	1995	no	
Journal of Development Economics[b] (JODE)	yes	2014	yes	2014
Journal of Econometrics (JOE)	no		no	
Journal of Environmental Economics and Management (JOEEM)	no		no	
Journal of Economic Surveys (JOES)	no		no	
Journal of Finance (JOF)	no		yes	2016
Journal of Financial Economics (JOFE)	no		no	
Journal of Financial Intermediation (JOFI)	no		no	
Journal of International Economics (JOIE)	no		no	
Journal of International Money and Finance (JOIMF)	no		no	
Journal of Labor Economics (JOLE)	yes	2009	yes	2009
Journal of Urban Economics (JOUE)	no		no	
Journal of Political Economy (JPE)	yes	2006	yes	2006
Journal of Public Economics (JPUE)	no		no	
Oxford Bulletin of Economics and Statistics (OBES)	no		no	
Quarterly Journal of Economics (QJE)	yes	2016	yes	2016
RAND Journal of Economics (RAND)	no		no	
Review of Economic Dynamics (RED)	no		no	
Review of Environmental Economics and Policy (REEP)	no		no	
Review of Economic Studies (RESTUD)	yes	2006	yes	2006
Review of Financial Studies (RFS)	yes	2020	yes	2020
World Bank Research Observer (WBRO)	no		no	

[a] Based on Christensen and Miguel (2018) the Journal of Money, Credit and Banking already implemented the policy in 1982 but discontinued the policy between 1993 and 1996.
[b] Based on the journal's website the starting month is August 2013. However, Askarov et al. (2022) state 2014 and we deem August late enough to count for 2014.
Source: Information obtained from journal websites and instructions for authors, as well as by email to journal staff through 2021 to 2022 and Askarov et al. (2022), Christensen and Miguel (2018), Müller-Langer et al. (2017), Vlaeminck and Herrmann (2015) and the Wayback Machine (cf. http://web.archive.org/, last retrieved on 20.03.2023).

Table B.6: Overview of all control variables

Variable name	Explanation	Characteristics
avg_len_table_notes	Average number of characters of table notes in whole article	positive float
contains_fixed_effects	Dummy if article contains regular expression of fixed effects	0 or 1
data_or_code_req	Dummy if test appears in a journal which either in this year or before implemented either a open data or open code policy	0 or 1
data_and_code_req	Dummy if test appears in a journal which either in this year or before implemented either a open data and open code policy	0 or 1
is_annex	Dummy if table is part of the annex/appendix	0 or 1
is_clustered	Dummy for usage of clustered standard errors in table	0 or 1
is_mult_testing	Dummy for usage multiple hypotheses testing in article	0 or 1
is_non-linear	Dummy for usage of certain non-linear models in article that may use the significance value of the underlying model instead	0 or 1
is_robustness	Dummy if table is robustness check	0 or 1
is_table_first_stage	Dummy for usage of first-stage regressions in table	0 or 1
no_authors	Number of authors of article	positive integer
no_authorsXY	Dummy if article has XY authors	0 or 1
no_tables_in_article	Number of tables in respective article	positive integer
no_tests_in_article	Number of tests in respective article	positive integer
no_tests_in_table	Number of tests in respective table	positive integer
number_strong_error_in_article	Number of any strong reporting error in respective article	positive integer
open_access	Dummy if article is open access	0 or 1
other_strong_error_in_article	Dummy if any other strong reporting error appears in respective article	0 or 1
pages_count	Number of pages of article	positive integer
ref_count	Number of references used in article	positive integer

Continued on next page

Table B.6 – *Continued from previous page*

Variable name	Explanation	Characteristics
sjr	Scimago Journal Rank of respective journal in respective year. 1998 imputed with 1999 for all journals available in 1999 and 2010 imputed with 2011 for *Annual Review of Economics* (ARE)	positive float
std_sig_levels	Dummy if standard significance levels (0.1, 0.05 and 0.01) were used in table	0 or 1
publisher_xy	Dummy if test is published in journal of publisher xy (xy $\in$ {Chicago, Elsevier, Oxford, Wiley}, Annual = default)	0 or 1
top5	Dummy if tests appears in one of the top 5	0 or 1
top_three_rows	Dummy if test is part of the top three rows in a table where tests appear	0 or 1
type_agg_p_value	Dummy if test type is of type p-value	0 or 1
type_agg_t_or_z_value	Dummy if test type is of type t-or z-value	0 or 1
type_no_se	Dummy if test type is not of type coefficient and standard error	0 or 1
volyear	Year of respective volume, normalized to 0 = 1998	positive integer

Table B.7: Overview of most important Regular Expressions

Topic (capture level)	Regular Expression	Comment
one-sided (article)	r'(?<=[\s\(>a])((?:left|right|one|1).(?:sided|tailed)|directional|side\stest)(?!ness)(?!\s(derivate|impact|aspect|violence|flow|trader|compound|limited\scommitment|selection|mass|inefficiency|repression|moving|gross\stransfer|imbalance))[^\w]'	Capture left-, right-, one-sided or -tailed as well as directional without capturing terms without a test context, e.g., one-sided commitment.
clustered standard errors (table)	r'(clustered|clustering|clusters?\s(over|at|by)|cluster\sstandard|allowing\s(for\s)?correlation\sat(?:\sthe)?\s[\w-]*?\slevel|standard\serrors(?:\sin\sparentheses,)?\scorrected\s(for|by)(?!\sheterosk)(?!\sautocorre)|standard\serrors(?:\sin\sparentheses,)?\sadjusted\sfor\s([\w-]*?\s)?cluster|cluster.robust|cluster\serrors\sby|intra\.-cluster\scorrelation|cluster.correct|obs\w+\sper\scluster|cluster\s(\([^()]*?\))?\sand\sheteroscedasticity\.-corrected\sstandard\serrors)'	Capture nouns describing clustered standard errors, but prevent capturing correction for heteroskedasticity and autocorrelation.
multiple testing (article)	r'(multiple\s(?:testing|comparison|p.value)|adjusted\saccording\sto\sbonferroni|bonferroni.s(correction|adjusted)|\sfwer\s|(?:family|experiment)wise\serror\srate|holm.bonferroni.method|[Šs]id[aá]k.correction|closed.testing.procedure|boole.bonferroni.bound|duncan(?:.s)?\snew\smultiple\srange\stest|harmonic\smean\sp.value\sprocedure|tukey(?:.s)?\srange\stest|hochberg(?:.s)?\sstep.up\sprocedure|dunnett(?:.s)?\stest|scheff[eé](?:.s)?\smethod|\sfdr\s|false\sdiscovery\srate|huynh[\s-]feld)'	Capture the most important methods of adjusting for multiple hypotheses testing as well as indicators, e.g., FDR.
robustness (table)	r'^table\s?[abcdef]?\.?[\div]{1,2}\.?\s?(?:robustness|extension|sensitivity|[\w\s(-\T1\textemdash]*?(?:robustness|sensitivity))'	Capture specific words in table heading (Note: \T1\textemdash is the character '—').
fixed-effects (article)	r'(fixed.effect|dumm(?:y|ies))'	Capture fixed-effects or dummies.

Continued on next page

Table B.7 – *Continued from previous page*

Topic (capture level)	**Regular Expression**	**Comment**
non-linear models (article)	r'(probit(?!y)|(?<!\=)logit|logistic(?!s|al)(?!\serror)|tobit|hazard\s(?:ratio|model)|poisson(?!\sshock)(?!\sprocess)(?!\sjump)|neg(?:ative)?\sbinomial)'	Capture non-linear models prone to report marginal effects but significance of underlying model.
appendix (table)	r'^(table\s?[abcdef]\.?\d|annex|appendix)'	Capture appendix and annex or special numbering in table headings but additionally, we check for position in article.

Table B.8: Prevalence of statistical reporting errors in main tests using different specifications

		One row			Two rows			Three rows			Four rows			Five rows		
Level	Type	O.	U.	A.	O.	U.	A.	O.	U.	A.	O.	U.	A.	O.	U.	A.
Journal	Any	83.33	86.67	90.00	90.00	93.33	93.33	93.33	93.33	93.33	93.33	93.33	93.33	93.33	93.33	93.33
	Strong	60.00	76.67	76.67	70.00	86.67	86.67	86.67	90.00	90.00	86.67	90.00	90.00	86.67	90.00	90.00
Article	Any	8.68	14.99	21.51	13.38	21.24	30.27	16.56	25.75	36.12	19.15	28.80	39.84	20.40	30.92	42.02
	Strong	3.78	3.62	7.21	6.58	5.49	11.45	8.43	7.51	14.88	9.95	9.11	17.57	10.80	10.61	19.34
Table	Any	2.91	5.27	7.95	4.78	8.30	12.54	6.16	10.53	15.81	7.19	12.04	18.04	7.91	13.31	19.71
	Strong	1.10	1.07	2.16	2.07	1.72	3.72	2.79	2.47	5.13	3.35	2.99	6.10	3.67	3.53	6.86
Test	Any	0.74	1.42	2.16	0.71	1.42	2.13	0.67	1.38	2.05	0.65	1.33	1.98	0.62	1.31	1.93
	Strong	0.27	0.26	0.53	0.28	0.23	0.51	0.26	0.24	0.51	0.26	0.24	0.50	0.24	0.25	0.49
No. of tests (articles)		65,810 (3,677)			122,005 (3,677)			172,040 (3,677)			216,421 (3,677)			255,405 (3,677)		
No. of tests (articles) afflicted with a strong reporting error		348 (265)			622 (421)			874 (547)			1,086 (646)			1,253 (711)		

Main tests are all tests that appear in the first x rows of a table but neither in a table labelled as robustness check nor a table in the appendix. O. = Overstated, U. = Understated, A. = Any.

Table B.9: Logistic regression at the test level with controls for strong overstated reporting errors

	All		Main tests		Non-main tests		First row	
Strong overstated								
(Intercept)	0.0022	0.0028	0.0028	0.0025	0.0020	0.0029	0.0027	0.0018
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
top5	1.1044	1.2441	0.8460	1.3668	1.2704	1.1785	1.0680	2.1725
	(0.6049)	(0.2914)	(0.5737)	(0.3812)	(0.3116)	(0.5173)	(0.8793)	(0.2039)
data_and_code_req	0.7401	0.8190	0.7815	0.8468	0.7111	0.8106	0.9532	0.9625
	(0.0109)	(0.0786)	(0.1179)	(0.3057)	(0.0214)	(0.1434)	(0.8431)	(0.8674)
sjr		1.0089		0.9982		1.0169		1.0008
		(0.4269)		(0.9114)		(0.2396)		(0.9752)
no_tests_in_article		0.9958		0.9960		0.9959		0.9966
		(0.0000)		(0.0000)		(0.0000)		(0.0044)
no_tables_in_article		1.0095		0.9859		1.0179		0.9822
		(0.4012)		(0.4575)		(0.1701)		(0.5713)
volyear		0.9821		1.0107		0.9634		1.0492
		(0.1258)		(0.5253)		(0.0117)		(0.0922)
no_authors2		1.0429		0.9323		1.1119		0.9514
		(0.6657)		(0.6508)		(0.3639)		(0.8359)
no_authors3		1.1002		0.9142		1.2318		1.1213
		(0.4077)		(0.5853)		(0.1531)		(0.6551)
no_authors4ormore		0.8942		0.6887		1.0625		0.7680
		(0.4976)		(0.1421)		(0.7586)		(0.4982)
avg_len_table_notes		1.0000		1.0004		0.9998		1.0005
		(0.6543)		(0.0024)		(0.1252)		(0.0178)
is_non_linear		1.1641		1.1889		1.1644		1.2431
		(0.0498)		(0.1456)		(0.1175)		(0.2492)
is_clustered		0.7718		0.7597		0.7715		0.7620
		(0.0049)		(0.0354)		(0.0189)		(0.1892)
is_annex		1.0652				1.1259		
		(0.6306)				(0.3943)		
is_robustness		0.7376				0.8060		
		(0.1040)				(0.2471)		
is_table_first_stage		1.3113		1.6467		1.0890		1.0169
		(0.0840)		(0.0162)		(0.6515)		(0.9753)
type_no_se		0.7530		0.8421		0.7291		0.8287
		(0.0174)		(0.3502)		(0.0129)		(0.4625)
other_strong_error_in_article		11.8205		13.2410		11.1677		13.5052
		(0.0000)		(0.0000)		(0.0000)		(0.0000)
pages_count		0.9936		0.9926		0.9932		0.9867
		(0.1488)		(0.2813)		(0.2119)		(0.2564)
ref_count		1.0007		0.9975		1.0027		0.9909
		(0.7798)		(0.5165)		(0.3964)		(0.1558)
std_sig_levels		0.7626		0.6909		0.8275		0.6067
		(0.0056)		(0.0127)		(0.1256)		(0.0499)
open_access		0.9107		0.9477		0.8960		0.9547
		(0.3463)		(0.7145)		(0.3528)		(0.8389)
McFadden Pseudo R^2	0.0009	0.1208	0.0009	0.1229	0.0011	0.1210	0.0000	0.1231
Num. obs.	578132	578132	172040	172040	406092	406092	65810	65810

Notes: Logistic regression with double lasso approach for variable selection of controls at the test level. Standard errors are clustered at the article level and based on 5,000 bootstrap replicates. Odds ratios with p-values in parentheses. Main tests refers to a subset of tests that appear in the first three rows of tables that are neither robustness checks nor appear in the appendix. First row refers to a subset of tests that appear in the first row of tables that are neither robustness checks nor appear in the appendix.

Table B.10: Logistic regression at the test level with controls for strong understated reporting errors

	All		Main tests		Non-main tests		First row	
Strong understated								
(Intercept)	0.0026	0.0014	0.0026	0.0009	0.0026	0.0016	0.0026	0.0024
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
top5	0.9507	1.1697	0.9917	1.3043	0.9393	1.1041	0.8394	1.7956
	(0.8059)	(0.4480)	(0.9787)	(0.4527)	(0.8050)	(0.7052)	(0.6992)	(0.2889)
data_and_code_req	0.8953	0.8794	0.7842	0.7810	0.9437	0.9200	0.9983	1.1322
	(0.3489)	(0.2195)	(0.1517)	(0.1442)	(0.6874)	(0.5227)	(0.9940)	(0.6211)
sjr		0.9876		0.9788		0.9917		0.9464
		(0.2707)		(0.2158)		(0.5699)		(0.1201)
no_tests_in_article		0.9975		0.9980		0.9973		0.9982
		(0.0000)		(0.0049)		(0.0000)		(0.0391)
no_tables_in_article		0.9859		0.9927		0.9863		1.0148
		(0.2456)		(0.7541)		(0.3529)		(0.6184)
volyear		1.0032		0.9792		1.0144		0.9383
		(0.7724)		(0.3285)		(0.2973)		(0.0635)
no_authors2		0.9472		0.9722		0.9431		1.1090
		(0.5680)		(0.8650)		(0.6043)		(0.6952)
no_authors3		0.9567		0.8658		0.9914		0.9199
		(0.6668)		(0.4619)		(0.9430)		(0.7559)
no_authors4ormore		0.8767		0.8485		0.9015		0.9205
		(0.3085)		(0.4889)		(0.5197)		(0.8078)
avg_len_table_notes		1.0000		0.9999		1.0000		1.0000
		(0.8335)		(0.3849)		(0.8113)		(0.9947)
is_non_linear		1.0840		0.9229		1.1316		0.7582
		(0.2616)		(0.5523)		(0.1417)		(0.1703)
is_clustered		1.2822		1.5881		1.1952		1.3063
		(0.0047)		(0.0010)		(0.0971)		(0.1670)
is_annex		1.1344				1.0476		
		(0.3280)				(0.7332)		
is_robustness		0.9437				0.8943		
		(0.7046)				(0.4768)		
is_table_first_stage		1.2591		1.2392		1.2882		1.1850
		(0.1368)		(0.4943)		(0.1378)		(0.6107)
type_no_se		0.9890		1.1756		0.9078		1.3078
		(0.9014)		(0.3588)		(0.3458)		(0.3732)
other_strong_error_in_article		15.8915		13.1352		17.1642		11.1965
		(0.0000)		(0.0000)		(0.0000)		(0.0000)
pages_count		1.0069		1.0129		1.0042		1.0072
		(0.1276)		(0.0537)		(0.4577)		(0.4823)
ref_count		0.9958		1.0003		0.9939		0.9943
		(0.1069)		(0.9413)		(0.0476)		(0.4315)
std_sig_levels		0.9062		1.4196		0.7573		1.4143
		(0.3321)		(0.0521)		(0.0211)		(0.2311)
open_access		0.9357		0.8589		0.9693		0.8804
		(0.4169)		(0.2792)		(0.7644)		(0.5697)
McFadden Pseudo R^2	0.0002	0.1229	0.0007	0.1094	0.0001	0.1322	0.0001	0.1127
Num. obs.	578132	578132	172040	172040	406092	406092	65810	65810

Notes: Logistic regression with double lasso approach for variable selection of controls at the test level. Standard errors are clustered at the article level and based on 5,000 bootstrap replicates. Odds ratios with p-values in parentheses. Main tests refers to a subset of tests that appear in the first three rows of tables that are neither robustness checks nor appear in the appendix. First row refers to a subset of tests that appear in the first row of tables that are neither robustness checks nor appear in the appendix.

Table B.11: Logistic regression at the test level without SJR

	All		Main tests		Non-main tests		First row	
Strong overstated								
Data and code required	0.7401	0.8019	0.7815	0.8508	0.7111	0.7814	0.9532	0.9603
	(0.0116)	(0.0486)	(0.1160)	(0.3063)	(0.0212)	(0.0849)	(0.8450)	(0.8640)
	[0.0464]	[0.1944]	[0.2998]	[0.4492]	[0.0848]	[0.3080]	[0.9940]	[0.8640]
Top 5	1.1044	1.3448	0.8460	1.3456	1.2704	1.3621	1.0680	2.1870
	(0.5990)	(0.0987)	(0.5740)	(0.3369)	(0.2960)	(0.1540)	(0.8778)	(0.1018)
	[0.7987]	[0.1974]	[0.7653]	[0.4492]	[0.5920]	[0.3080]	[0.9940]	[0.4072]
Controls	No	Yes	No	Yes	No	Yes	No	Yes
McFadden's Pseudo R^2	0.0009	0.1208	0.0009	0.1229	0.0011	0.1208	0.0000	0.1231
Strong understated								
Data and code required	0.8953	0.9066	0.7842	0.8261	0.9437	0.9381	0.9983	1.2705
	(0.3524)	(0.3350)	(0.1499)	(0.2592)	(0.6822)	(0.6068)	(0.9940)	(0.3375)
	[0.7048]	[0.4467]	[0.2998]	[0.4492]	[0.8031]	[0.8091]	[0.9940]	[0.6750]
Top 5	0.9507	1.0446	0.9917	1.0626	0.9393	1.0268	0.8394	1.0935
	(0.8049)	(0.8039)	(0.9786)	(0.8499)	(0.8031)	(0.9096)	(0.6707)	(0.8543)
	[0.8049]	[0.8039]	[0.9786]	[0.8499]	[0.8031]	[0.9096]	[0.9940]	[0.8640]
Controls	No	Yes	No	Yes	No	Yes	No	Yes
McFadden's Pseudo R^2	0.0002	0.1228	0.0007	0.1090	0.0001	0.1322	0.0001	0.1103
Observations	578,132	578,132	172,040	172,040	406,092	406,092	65,810	65,810

Notes: Logistic regression with double lasso approach for variable selection of controls at the test level. Standard errors are clustered at the article level and based on 5,000 bootstrap replicates. Odds ratios with p-values in parentheses and FDR-adjusted p-values in brackets depicted. Intercept not reported. Information on the control variables is given in Table B.6 in the Online Appendix. In this regression SJR is not part of the control variables. Main tests refers to a subset of tests that appear in the first three rows of tables that are neither robustness checks nor appear in the appendix. First row refers to a subset of tests that appear in the first row of tables that are neither robustness checks nor appear in the appendix.

Table B.12: Logistic regression at the article level

	All		Main tests		Non-main tests		First row	
Strong overstated								
Data and code required	0.8026	0.9541	0.8566	1.0890	0.7054	0.8141	0.8669	1.0499
	(0.0502)	(0.7186)	(0.3014)	(0.6321)	(0.0078)	(0.2065)	(0.5164)	(0.8387)
	[0.2008]	[0.7186]	[0.8082]	[0.6639]	[0.0312]	[0.6012]	[0.9335]	[0.9412]
Top 5	0.9769	0.9084	0.8896	0.8415	1.0228	0.9448	1.2060	1.2988
	(0.8995)	(0.6931)	(0.6610)	(0.6234)	(0.9196)	(0.8421)	(0.5937)	(0.6112)
	[0.8995]	[0.7186]	[0.8813]	[0.6639]	[0.9196]	[0.8421]	[0.9335]	[0.9412]
Controls	No	Yes	No	Yes	No	Yes	No	Yes
McFadden's Pseudo R^2	0.0012	0.0415	0.0007	0.0398	0.0026	0.0499	0.0005	0.0401
Strong understated								
Data and code required	0.8526	0.9081	0.8730	0.8717	0.8395	0.9152	1.0800	1.0684
	(0.1346)	(0.4534)	(0.4041)	(0.4754)	(0.1500)	(0.5491)	(0.7250)	(0.8024)
	[0.2692]	[0.7186]	[0.8082]	[0.6639]	[0.2000]	[0.7321]	[0.9335]	[0.9412]
Top 5	0.7849	0.8298	1.0100	0.8546	0.6683	0.7427	1.0337	1.0402
	(0.2104)	(0.4452)	(0.9717)	(0.6639)	(0.0850)	(0.3006)	(0.9335)	(0.9412)
	[0.2805]	[0.7186]	[0.9717]	[0.6639]	[0.1700]	[0.6012]	[0.9335]	[0.9412]
Controls	No	Yes	No	Yes	No	Yes	No	Yes
McFadden's Pseudo R^2	0.0014	0.0517	0.0004	0.0584	0.0023	0.0543	0.0001	0.0642
Observations	3,746	3,746	3,677	3,677	3,611	3,611	3,677	3,677

Notes: Logistic regression with double lasso approach for variable selection of controls at the article level. The dependent variable is a dummy variable that equals 1 if the article contains at least one strong overstated reporting error or one strong understated reporting error respectively. Standard errors are based on 5,000 bootstrap replicates. Odds ratios with p-values in parentheses and FDR-adjusted p-values in brackets depicted. Intercept not reported. Information on the control variables is given in Table B.6 in the Online Appendix. Main tests refers to a subset of tests that appear in the first three rows of tables that are neither robustness checks nor appear in the appendix. First row refers to a subset of tests that appear in the first row of tables that are neither robustness checks nor appear in the appendix.

Table B.13: Number of articles per biannual period per journal

	1999-2000	2001-2002	2003-2004	2005-2006	2007-2008	2009-2010	2011-2012	2013-2014	2015-2016	Sum
ARE	0	0	0	0	0	1	0	1	2	4
ARFE	0	0	0	0	0	0	0	0	1	1
EER	3	6	11	16	19	13	31	44	86	229
EJ	0	0	11	16	28	41	26	37	65	224
EP	0	4	4	5	14	13	14	9	13	76
IER	0	0	0	0	2	3	6	5	9	25
JEEA	0	0	0	0	0	0	8	17	23	48
JMCB	0	0	0	0	14	33	12	20	16	95
JME	1	2	8	8	5	7	3	10	12	56
JOAE	0	3	4	3	4	2	9	6	4	35
JODE	6	13	13	23	38	45	70	102	83	393
JOE	1	0	1	0	6	3	3	4	5	23
JOEEM	0	0	10	5	8	14	22	36	19	114
JOES	0	0	0	0	0	1	4	1	2	8
JOF	17	15	13	30	28	35	40	43	61	282
JOFE	6	12	19	23	47	50	68	78	89	392
JOFI	0	0	2	2	1	7	15	18	25	70
JOIE	2	4	11	18	21	23	44	51	74	248
JOIMF	2	1	4	9	13	26	33	56	61	205
JOLE	0	12	9	12	12	15	13	26	35	134
JOUE	0	0	0	0	0	20	27	37	34	118
JPE	3	7	11	10	10	15	13	12	17	98
JPUE	6	9	16	29	37	47	55	81	75	355
OBES	0	1	6	3	9	8	14	10	17	68
QJE	0	0	0	0	0	0	26	31	29	86
RAND	0	0	0	0	5	3	5	12	7	32
RED	0	0	0	0	0	0	0	2	3	5
REEP	0	0	0	0	0	0	1	0	0	1
RESTUD	0	0	0	0	0	0	13	20	23	56
RFS	6	4	7	4	13	66	47	46	52	245
WBRO	0	0	0	0	0	0	1	0	0	1
Sum	53	93	160	216	334	491	623	815	942	3,727

Notes: The highlighted rows indicate the journals that have at least 100 articles in total, for which DORIS could detect at least 10 articles biannually (exception JOLE), and that either never implemented an open data and code policy or have at least three time periods since 2003 in which they had no open data and code policy (hence, no JPE). A list of abbreviations for the journals can be found in Table B.5 in the Online Appendix.

Table B.14: BJS imputation estimator at the test level: Falsification test

	All		Main tests		Non-main tests		First row	
No anticipation								
Strong overstated								
Data and code required	-0.0001	-0.0002	-0.0005	-0.0009	0.0000	0.0000	-0.0010	-0.0015
p-value	(0.9359)	(0.8715)	(0.7395)	(0.5527)	(0.9734)	(0.9840)	(0.6919)	(0.5454)
Pre-trend test	[0.1463]	[0.3592]	[0.9838]	[0.9236]	[0.0703]	[0.1369]	[0.4859]	[0.3661]
Strong understated								
Data and code required	0.0000	0.0002	0.0015	0.0015	-0.0007	-0.0004	0.0029	0.0029
p-value	(0.9708)	(0.8441)	(0.2510)	(0.2497)	(0.6101)	(0.7459)	(0.1703)	(0.1442)
Pre-trend test	[0.1100]	[0.2301]	[0.0584]	[0.2742]	[0.1920]	[0.4150]	[0.9989]	[0.4811]
Journal and year effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	249,126	249,126	74,242	74,242	174,884	174,884	28,438	28,438

Notes: Falsification test of BJS imputation estimator at the test level of Table 5 showing the average treatment effect on the treated as probabilities. The falsification is carried out by assuming that the respective policy commenced four years (i.e. two biannuals) earlier. The dependent variable in the first half of the table is a dummy, that is, if a test is afflicted with a strong overstated reporting error. The dependent variable in the second half of the table is a dummy, that is, one if a test is afflicted with a strong understated reporting error. Standard errors are clustered at the article level. p-values of the coefficients in parentheses and p-values of pre-trend tests with three periods in brackets depicted. Controls include dummy variables for the number of authors, the number of tests per article, dummies for the test type, for the usage of standard significance levels, as well as for the prevalence of clustered standard errors in the corresponding table and the occurrence of another strong reporting error within the same article. The data set comprises the journals EER, EJ, JODE, JOF, JOFE, JOIE, JOLE and JPUE from 2003 to 2016 and defines every two years as one time period. Main tests refers to a subset of tests that appear in the first three rows of tables that are neither robustness checks nor appear in the appendix. First row refers to a subset of tests that appear in the first row of tables that are neither robustness checks nor appear in the appendix.

C Figures

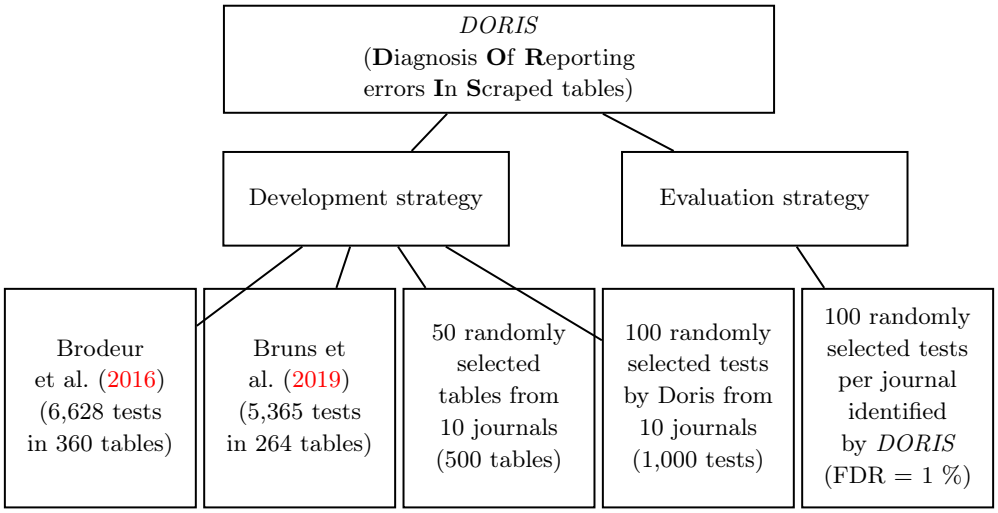


Figure C.4: Development and evaluation strategy, own illustration

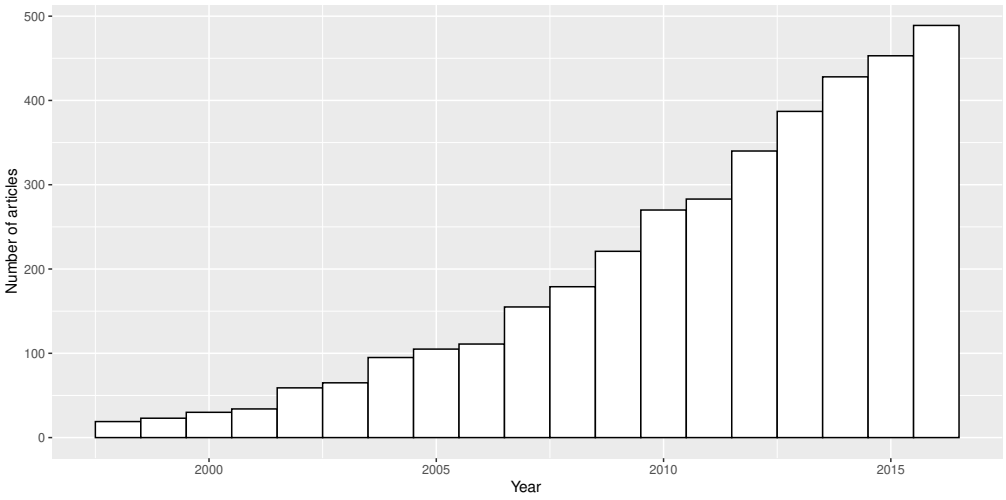


Figure C.5: Articles per year in the manually corrected sample without outliers

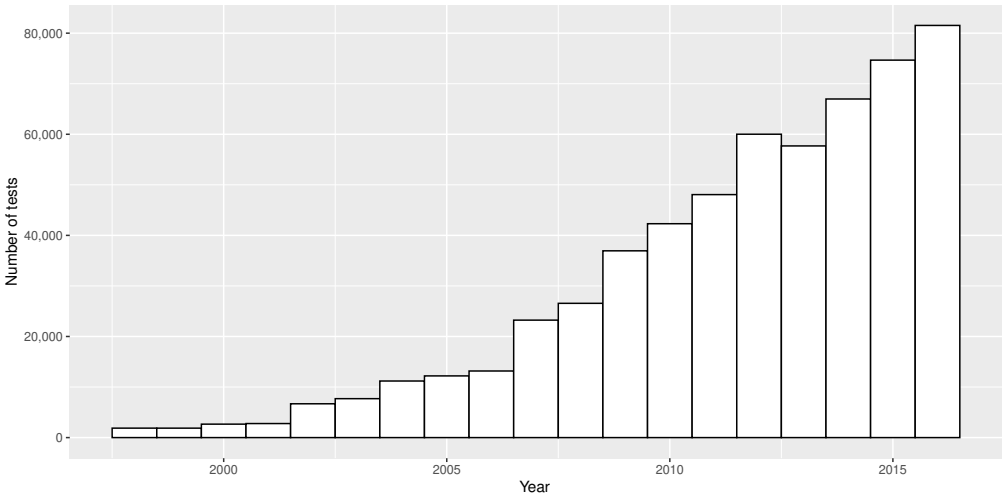


Figure C.6: Tests per year in the manually corrected sample without outliers

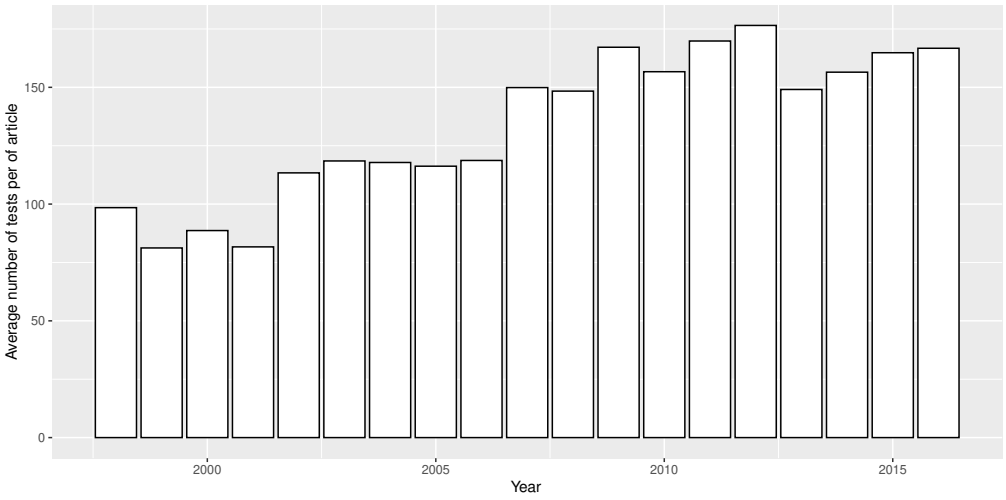


Figure C.7: Tests per article per year in the manually corrected sample without outlier

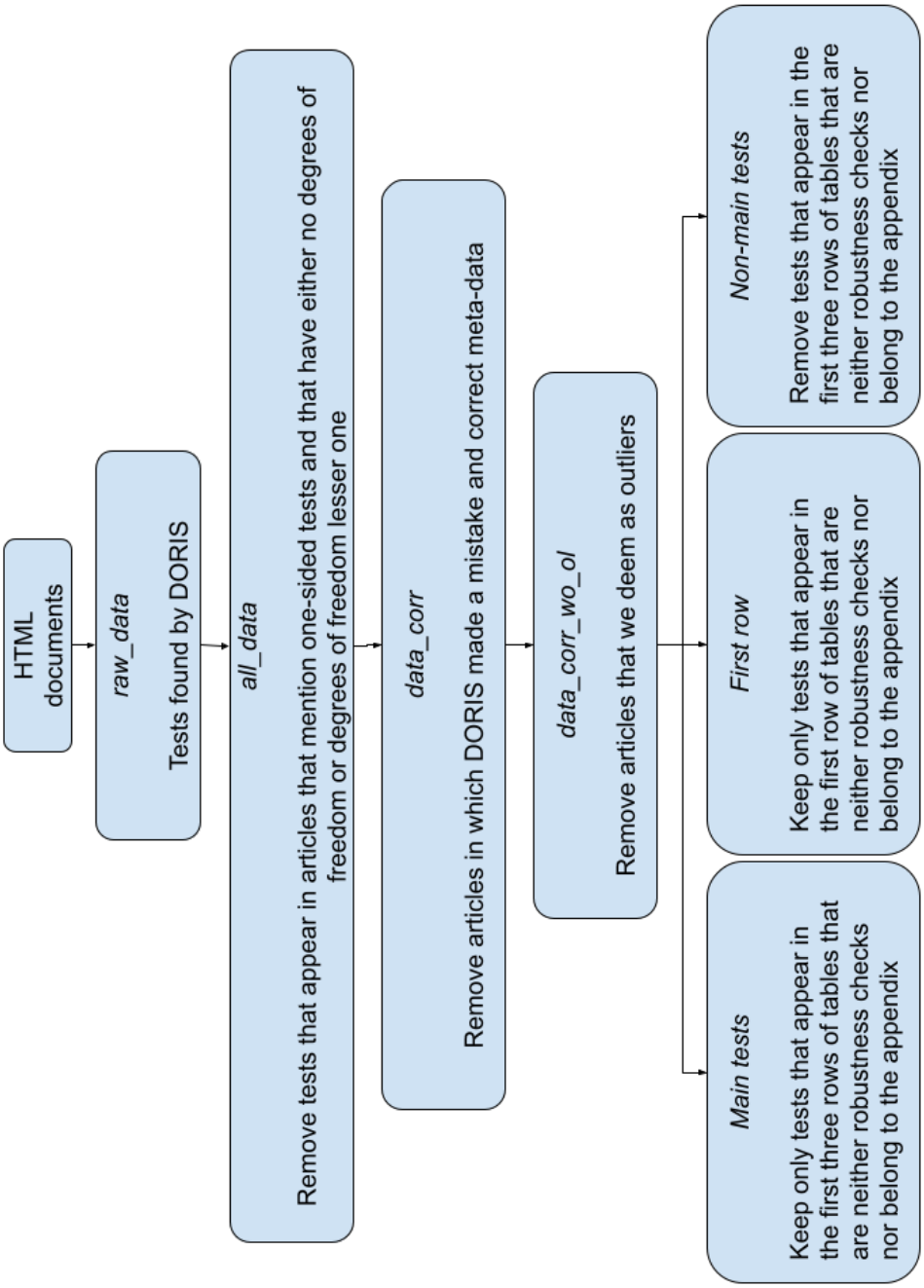


Figure C.8: Overview of used data sets

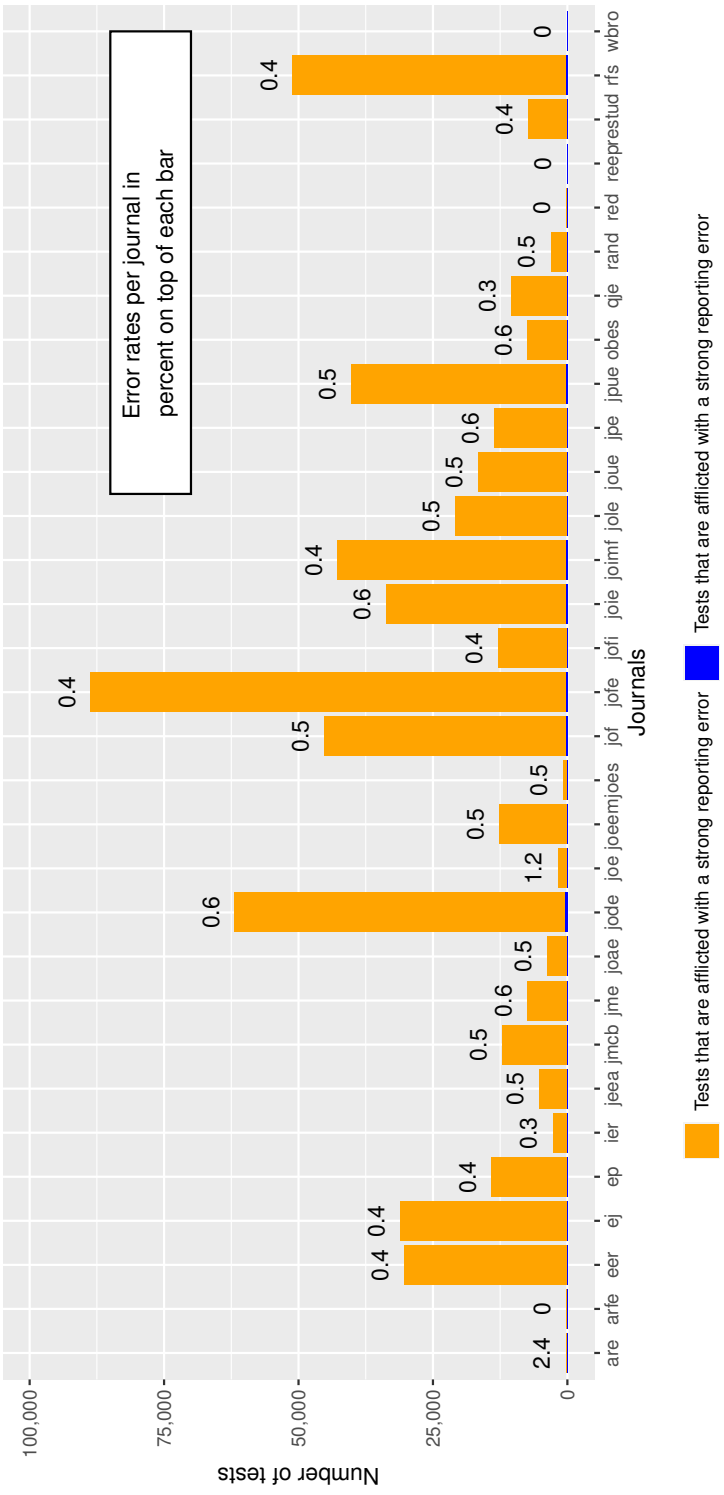
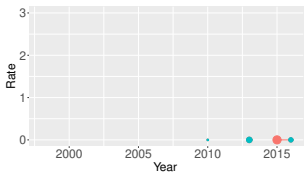
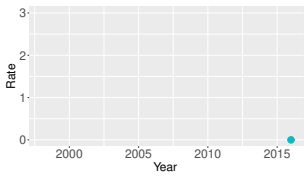


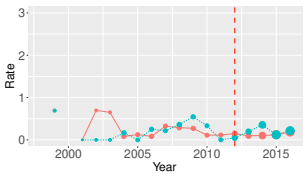
Figure C.9: Tests and strong reporting errors per journal considering all tests



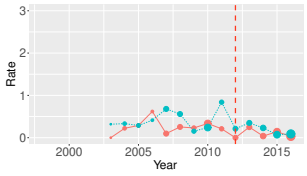
(a) Annual Review of Economics



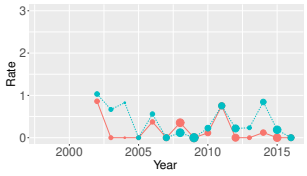
(b) Annual Review of Financial Economics



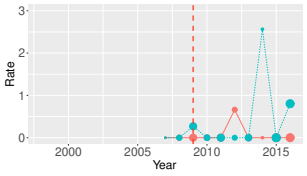
(c) European Economic Review



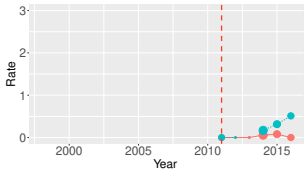
(d) Economic Journal



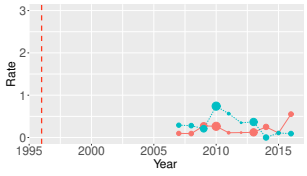
(e) Economic Policy



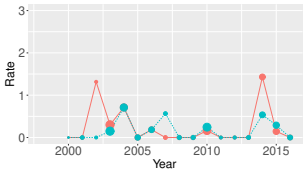
(f) International Economic Review



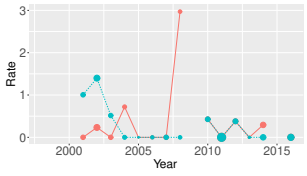
(g) Journal of the European Economic Association



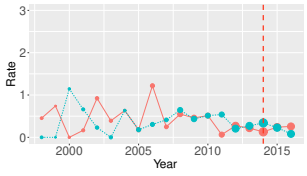
(h) Journal of Business & Economic Statistics



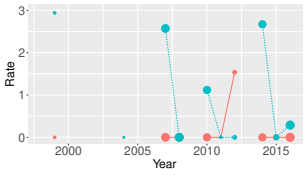
(i) Journal of Monetary Economics



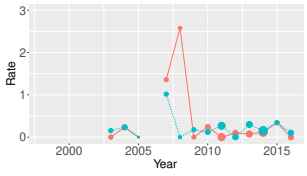
(j) Journal of Applied Econometrics



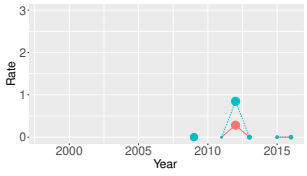
(k) Journal of Development Economics



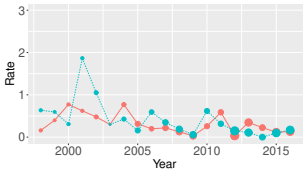
(l) Journal of Econometrics



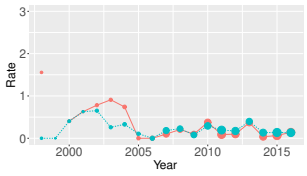
(m) Journal of Environmental Economics and Management



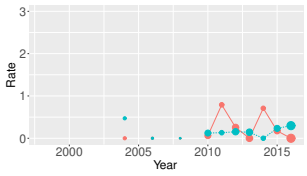
(n) Journal of Economic Surveys



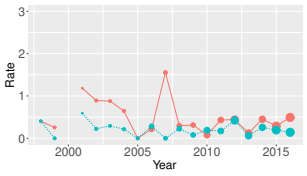
(o) Journal of Finance



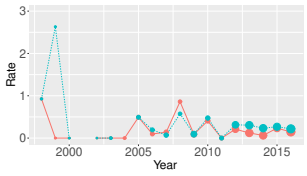
(p) Journal of Financial Economics



(q) Journal of Financial Intermediation

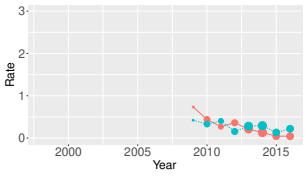


(r) Journal of International Economics

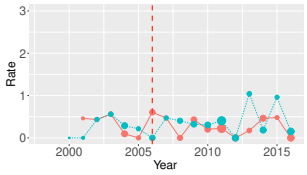


(s) Journal of International Money and Finance

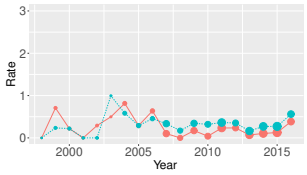
(t) Journal of Labor Economics



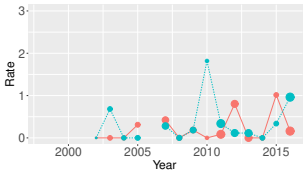
(u) Journal of Urban Economics



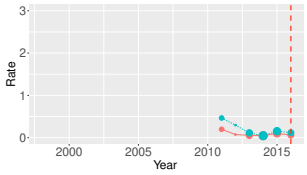
(v) Journal of Political Economy



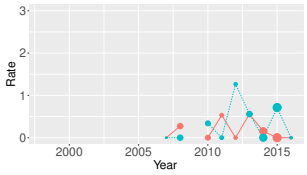
(w) Journal of Public Economics



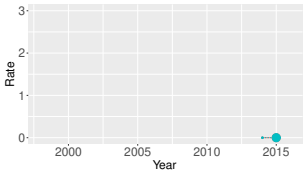
(x) Oxford Bulletin of Economics and Statistics



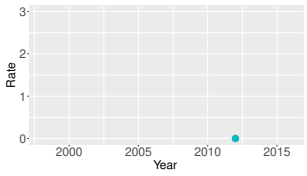
(y) Quarterly Journal of Economics



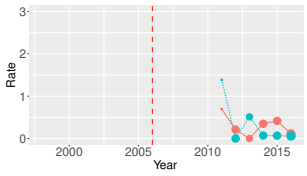
(z) RAND Journal of Economics



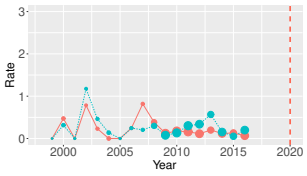
(aa) Review of Economic Dynamics



(ab) Review of Environmental Economics and Policy



(ac) Review of Economic Studies

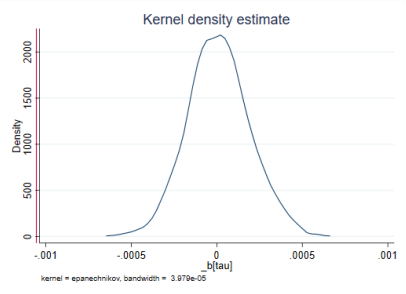


(ad) Review of Financial Studies

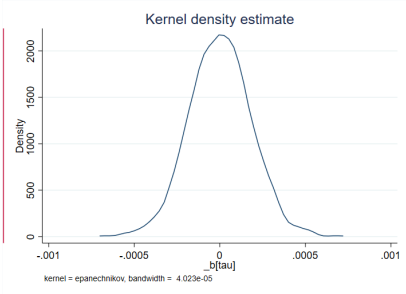
(ae) World Bank Research Observer (af) All journals

Notes: Rate of strong overstated reporting errors among all tests of the respective journal and year as well as rate of strong understated reporting errors among all tests of the respective journal and year. The vertical dashed line represents the year of implementation of a mandatory data and code policy. The larger a dot, the more tests are in the respective sample. Sizes are not comparable across subfigures.

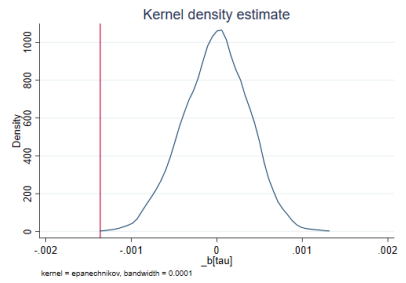
Figure C.10: Reporting error rates for strong over- and understated reporting errors per journal over time (all tests)



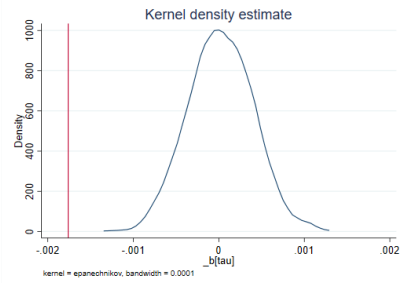
(a) All tests w/o controls



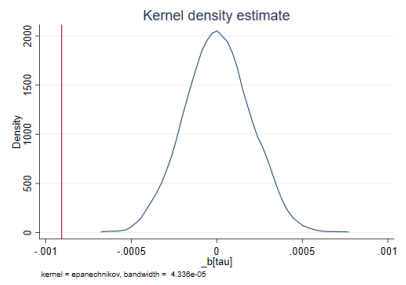
(b) All tests with controls



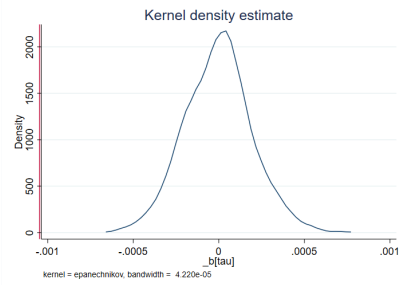
(c) Main tests w/o controls



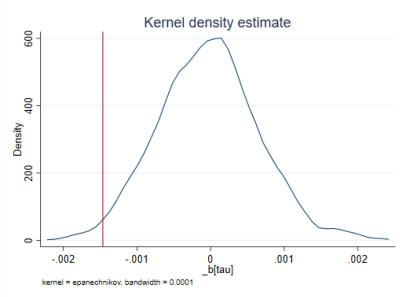
(d) Main tests with controls



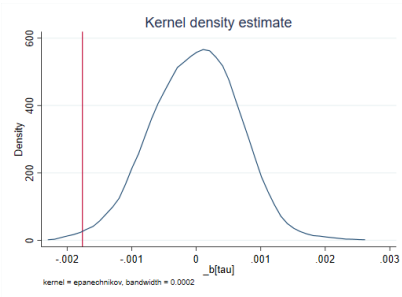
(e) Non-main tests w/o controls



(f) Non-main tests with controls



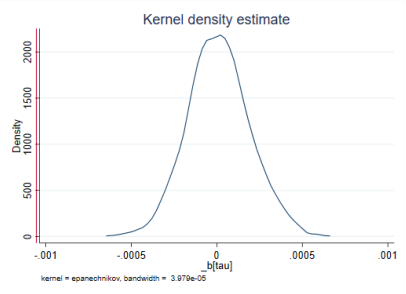
(g) First row w/o controls



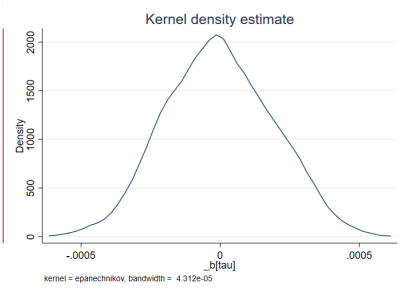
(h) First row tests with controls

Notes: Falsification test of BJS imputation estimator of Table 5 for strong overstated reporting errors as dependent variable at the test level showing the kernel density estimates. The falsification test is performed by permuting the treatment variable 1,000 times. The red line indicates the ATT of the main model. An accumulation of the density around zero and a red line close to the margins indicates a robust finding. No anticipation is assumed. All estimates contain journal and year fixed effects. Controls include dummy variables for the number of authors, the number of tests per article, dummies for the test type, for the usage of standard significance levels as well as for the prevalence of clustered standard errors in the corresponding table and the occurrence of another strong reporting error within the same article. The data set comprises the journals EER, EJ, JODE, JOF, JOFE, JOIE, JOLE and JPUE from 2003 to 2016 and defines every two years as one time period, i.e. one year. Main tests refers to a subset of tests that appear in the first three rows of tables that are neither robustness checks nor appear in the appendix. First row refers to a subset of tests that appear in the first row of tables that are neither robustness checks nor appear in the appendix.

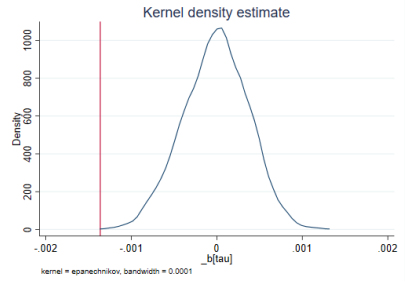
Figure C.11: Falsification test for BJS imputation estimator at the test level for strong overstated reporting errors



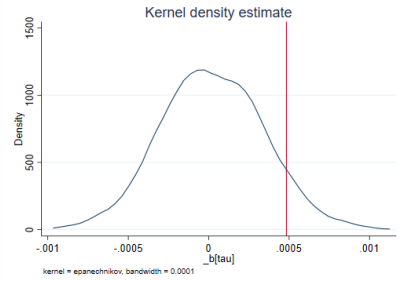
(a) All tests w/o controls



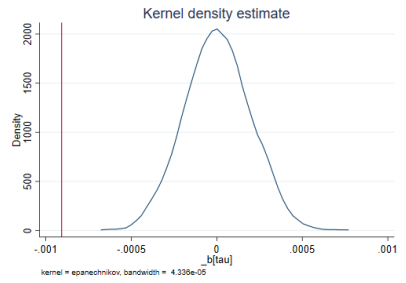
(b) All tests with controls



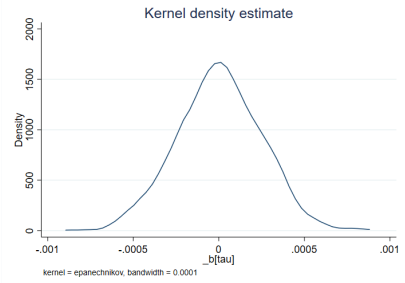
(c) Main tests w/o controls



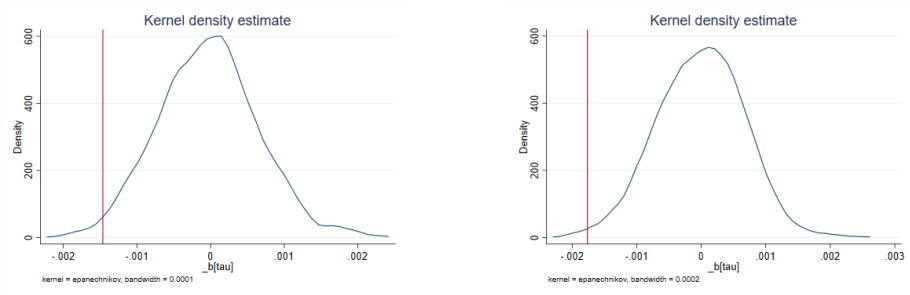
(d) Main tests with controls



(e) Non-main tests w/o controls



(f) Non-main tests with controls



(g) First row w/o controls (h) First row tests with controls

Notes: Falsification test of BJS imputation estimator of Table 5 for strong understated reporting errors as dependent variable at the test level showing the kernel density estimates. The falsification test is performed by permuting the treatment variable 1,000 times. The red line indicates the ATT of the main model. An accumulation of the density around zero and a red line close to the margins indicates a robust finding. No anticipation is assumed. All estimates contain journal and year fixed effects. Controls include dummy variables for the number of authors, the number of tests per article, dummies for the test type, for the usage of standard significance levels as well as for the prevalence of clustered standard errors in the corresponding table and the occurrence of another strong reporting error within the same article. The data set comprises the journals EER, EJ, JODE, JOF, JOFE, JOIE, JOLE and JPUE from 2003 to 2016 and defines every two years as one time period, i.e. one year. Main tests refers to a subset of tests that appear in the first three rows of tables that are neither robustness checks nor appear in the appendix. First row refers to a subset of tests that appear in the first row of tables that are neither robustness checks nor appear in the appendix.

Figure C.12: Falsification test for BJS imputation estimator at the test level for strong understated reporting errors