Online Appendix

Descriptive statistics

The distributions of (strong) reporting errors per article after taking into account the survey responses and the replication results are given in Table A1 and Table A2.

Table A1: Distribution of reporting errors per article

Number of errors	0	1	2	3	4	5	6	7	8	9	10	11	16	18	19
Frequency	208	68	29	24	16	7	4	2	1	3	3	2	1	1	1

Table A2: Distribution of strong reporting errors per article

Number of errors	0	1	2	3	4	5	15
Frequency	272	63	21	7	4	2	1

Definitions of the variables used in the regression analyses in Section 6 of the article are presented below and further descriptive information is given in Tables A3 and A4.

- Journal id: The journal the article was published in.
- Field: Whether the article is from microeconomics or macroeconomics.
- Negative results: Whether the article presents at least one null result in contrast to only positive results as contribution.
- Theoretical model: Whether the article includes a theoretical contribution.
- Data availability: Whether the data of the article is available on the respective website of the journal.
- Code availability: Whether the software code of the article is available on the respective website of the journal.
- Year: The publication year of the article.
- Number of authors: The number of authors of the article.
- Share of editors among the authors: Share of authors who are member of an editorial board at the time of publication.
- Share of tenured authors: Share of authors who are full professor at the time of publication.
- Authors' average years since PhD: Authors' average years since PhD at the time of publication.
- Number of research assistants thanked: Number of research assistants thanked in the article.
- Number of individuals thanked: Number of individuals thanked excluding research assistants and referees.

- Number of tables: Number of tables reporting results of hypothesis tests that test central hypotheses; tables with several panels are treated as one table.

- Number of tests: Number of hypothesis tests that test central hypotheses.

Details on the original reporting guidelines for the variables can be found in the Online Appendix of Brodeur et al. (2016).

Table A3: Distribution of discrete variables at the test and article levels

	Number of Articles	Number of Tests
American Economic Review	180	13247
Journal of Political Economy	61	4997
Quarterly Journal of Economics	129	12749
Macroeconomics	85	8142
Microeconomics	285	22851
Negative results: yes	60	5848
Negative results: no	310	25145
With theoretical model	109	8964
Without theoretical model	261	22029
Data available	177	13553
Data not available	193	17440
Code available	195	15214
Code not available	175	15779

Table A4: Descriptive statistics for the continuous variables at the article level

Variable	n	Min	Q1	Median	Mean	Q3	Max	SD
Year	370	1	3.0	4.4	5.0	6.0	7	2.0
Number of authors	370	1	2.0	2.2	2.0	3.0	5	0.9
Share of editors among authors	370	0	0.0	0.4	0.3	0.6	1	0.4
Share of tenured authors	370	0	0.0	0.3	0.3	0.5	1	0.3
Authors' average years since PhD	367	-2	5.5	10.1	9.0	12.8	41	6.6
Number of research assistants thanked	370	0	0.0	2.2	1.0	3.0	27	3.5
Number of individuals thanked	370	0	6.0	11.2	9.5	15.0	45	7.7
Number of tables	370	1	2.0	4.1	4.0	5.0	15	2.4
Number of tests	370	2	30.2	83.8	60.0	107.8	587	82.9

Propensity to respond to the survey

We run logistic regressions to analyze whether replying to our survey is associated to observable characteristics regarding the authors and the articles. Table A5 indicates that the probability to answer increases as the number of tests conducted in the article grows. All in all, the explanatory power of the regression models are very low (Pseudo R^2 of below 5%). As indicated by the AIC, the enhanced complexity of the two models including several explanatory variables outweigh the low increase in the explanatory power compared to the intercept-only model which performs best in this regard.

Robustness checks

Our replication based estimate is that 100% of the flagged tests that imply an overstated significance level are correctly flagged. We explore the robustness of this finding using a variable gathered by Brodeur et al. (2016) indicating the type of statistical model/test used in the respective articles. Non-standard reporting was the main reason for falsely flagged tests with overstated significance levels and non-standard reporting seems to occur in association with non-linear models. According to the authors, a non-standard reporting style was identified as the reason for a falsely flagged test in 45 cases. In 41 of these 45 cases, the variable indicating the type of statistical model shows that a logit, probit or a model different to linear regression was applied. Among the 285 flagged tests in the replication sample, the same holds for only twelve overstated significance levels. Of these twelve flagged tests, seven were indeed reporting

Table A5: Propensity to respond to the survey – regression results

	Response to survey	Response to survey	Response to survey
Intercept	-0.982	-104.076	-85.859
	[-1.184; -0.810]	[-330.470; 112.303]	[-316.276; 138.335]
Year		0.051	0.042
		[-0.056; 0.164]	[-0.069; 0.157]
Journal of Political Economy		0.330	0.190
		[-0.272; 0.917]	[-0.554; 0.877]
Quarterly Journal of Economics		-0.251	-0.472
		[-0.742; 0.291]	[-1.420; 0.469]
Field: Macroeconomics		-0.236	-0.258
		[-0.773; 0.300]	[-0.795; 0.273]
No. of authors		-0.194	-0.211
		[-0.426; 0.071]	[-0.448; 0.059]
Share of editors among authors		-0.270	-0.327
		[-0.928; 0.432]	[-0.980; 0.381]
Share of tenured authors		0.212	0.323
		[-0.817; 1.076]	[-0.760; 1.177]
Authors' average years since PhD		-0.001	-0.008
		[-0.053; 0.049]	[-0.059; 0.043]
No. of research assistants thanked		0.023	0.026
		[-0.039; 0.085]	[-0.041; 0.086]
No. of individuals thanked		0.004	0.005
		[-0.022; 0.034]	[-0.022; 0.035]
Negative results put forward		0.145	0.146
		[-0.463; 0.751]	[-0.460; 0.758]
With theoretical model		-0.233	-0.180
		[-0.729; 0.273]	[-0.696; 0.325]
No. of tables		-0.027	-0.014
		[-0.137; 0.084]	[-0.126; 0.100]
No. of tests		0.005	0.005
		[0.002; 0.007]	[0.002; 0.007]
Data available			0.733
			[-0.091; 1.685]
Code available			-0.910
			[-1.787; 0.043]
n	367	367	367
Pseudo R^2	0.000	0.038	0.048
AIC	431.911	443.503	443.142

Notes: Results from logistic regressions are shown. The dependent variable is a dummy variable indicating whether an author responded to our survey. Lower and upper bounds of 90% bias corrected and accelerated (BCa) intervals based on 5000 bootstrap replicates in brackets.

errors as shown by the replications and five were not replicated. We thus conclude that the estimated share of 100% correctly flagged tests with overstated significance levels appears to be reliable.

In the following, we furthermore describe two robustness checks for estimating the rates of reporting errors. The robustness checks in Table A6 should be compared with column three in Table 3 in the main body of the paper.

The share of zeros as last reported decimal is only about 5.6%, whereas each of the other digits accounts for 9.8-11.2%. A manual check confirmed that zeros are occasionally missing as last decimals in the data of Brodeur et al. (2016). Since it was infeasible for us to check all reported numbers manually, we approached the effect of dropping zeros on our estimates by rerunning our analyses after dropping the last decimals if they are equal to zero. We found almost no difference in the results (Table A6, column one).

As a further robustness check, we examined whether potentially trimmed decimals affect our estimates, e.g. a reported coefficient of 1.4 was originally 1.48. To this end, we dropped all last decimals. While emphasizing that such a trimming procedure involves erroneous rounding, namely rounding down when rounding up would be adequate, the estimated rates change only marginally (Table A6, column two).

Table A6: Prevalence of reporting errors - robustness checks

			Zeros removed	Trimmed decimals
Article level	Any error	Overstated	96 (25.95%)	95 (25.68%)
		Understated	90 (24.22%)	86 (23.12%)
		Any	129 (34.95%)	126 (33.97%)
	Strong error	Overstated	52 (14.05%)	51 (13.78%)
		Understated	44 (11.82%)	42 (11.31%)
		Any	79 (21.34%)	75 (20.21%)
Test level	Any error	Overstated	206 (0.66%)	200 (0.65%)
		Understated	193 (0.62%)	179 (0.58%)
		Sum	399 (1.29%)	379 (1.22%)
	Strong error	Overstated	80 (0.26%)	76 (0.25%)
		Understated	69 (0.22%)	64 (0.21%)
		Sum	149 (0.48%)	140 (0.45%)

Notes: Numbers and shares of any and strong reporting errors at the test and article level are given. "Overstated" means overstated significance level compared to the calculated p-value, "Understated" means understated significance level compared to the calculated p-value. The estimates are calculated after including information from the survey responses and the replications. The absolute numbers are rounded estimates, see Section **??** for the corresponding estimation strategy. The first column shows error rates calculated under the condition that the last decimals of the reported statistical values are removed if they are equal to zero. The second column shows error rates calculated under the condition that authors trimmed the reported statistical values instead of rounding properly.

References

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