Star Wars: The Empirics Strike Back

Abel Brodeur, Mathias Lé, Marc Sangnier, Yanos Zylberberg

amse
école d'économie d'aix-marseille
aix-marseille school of economics

Working Papers / Documents de travail

Star Wars: The Empirics Strike Back

Abel Brodeur
Mathias Lé
Marc Sangnier
Yanos Zylberberg

WP 2015 - Nr 23

Star Wars: The Empirics Strike Back*

Abel Brodeur Mathias Lé Marc Sangnier
Yanos Zylberberg

May 2015

Abstract

Journals favor rejection of the null hypothesis. This selection upon tests may distort the behavior of researchers. Using 50,000 tests published between 2005 and 2011 in the *AER*, *JPE*, and *QJE*, we identify a residual in the distribution of tests that cannot be explained by selection. The distribution of p-values exhibits a two humped camel shape with abundant p-values above 0.25, a valley between 0.25 and 0.10, and a bump slightly below 0.05. The missing tests (with p-values between 0.25 and 0.10) can be retrieved just after the 0.05 threshold and represent 10% to 20% of marginally rejected tests. Our interpretation is that researchers might be tempted to *inflate* the value of those just-rejected tests by choosing a "significant" specification. We propose a method to measure this residual and describe how it varies by article and author characteristics.

KEYWORDS: Hypothesis testing, distorting incentives, selection bias, research in economics.

JEL CODES: A11, B41, C13, C44.

*Brodeur: University of Ottawa; Lé: Paris School of Economics and ACPR; Sangnier: Aix-Marseille Univ. (Aix-Marseille School of Economics), CNRS & EHESS; Zylberberg (corresponding author): School of Economics Finance and Management, University of Bristol; Priory Road Complex, BS81TU Bristol, +44 117 92 88419, yanos.zylberberg@bristol.ac.uk. We thank Orley Ashenfelter, Regis Barnichon, Marianne Bertrand, Thomas Breda, Paula Bustos, Colin Camerer, Andrew Clark, Gabrielle Fack, Jordi Gali, Nicola Gennaioli, Alan Gerber, Libertad González, David Hendry, Emeric Henry, James MacKinnon, Ted Miguel, Steve Pischke, Thijs van Rens, Tom Stanley, and Rainer Winkelmann, seminar participants at AMSE, CREI, and Universitat Pompeu Fabra, and the audiences at AFSE 2012, EEA 2012, LAGV 2013, MAER-Net Colloquium 2013, and Workshop on the Economics of Science 2013 for very useful remarks and encouragements. Financial support from the FQRSC, SSHRC, Région Île-de-France, and the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement n° 241114 is gratefully acknowledged by Abel Brodeur, Mathias Lé, and Yanos Zylberberg, respectively. Financial support from PSE Research Fund is collectively acknowledged by the authors.

If the stars were mine
I'd give them all to you
I'd pluck them down right from the sky
And leave it only blue.
"If The Stars Were Mine" by Melody Gardot

1 Introduction

The introduction of norms—confidence at 95% or 90%—and the use of eye-catchers—stars—have led the academic community to accept more easily starry stories with marginally significant coefficients than starless ones with marginally insignificant coefficients.[1] As highlighted by Sterling (1959), this effect has modified the selection of papers published in journals and arguably biased publications toward tests rejecting the null hypothesis. This selection is not unreasonable. The choice of a norm was precisely made to strongly discriminate between rejected and accepted hypotheses.

A consequence of such selection is that researchers may anticipate and consider that it is a stumbling block for their ideas to be considered. For instance, they may censor their papers with too high p-values. They may also search for specifications delivering just-significant results and ignore specifications giving just-insignificant results in order to increase their chances of being published.[2] The latter behavior has different implications on the distribution of published tests than selection by journals. If selection by journals is monotonically increasing with the value of test statistics, the proportion of submitted papers that end up being published should increase with the value of test statistics. We build on this assumption to propose an accounting framework that can be applied to any distribution of published test statistics. This method allows us to determine what could be attributed to *selection* in the distribution of published statistics and extract a residual that cannot be explained by this selection process alone. We thereafter refer to this residual as *inflation* because it should capture, among other things, part of the behavioral response.

Why would we expect the distribution of published statistics to be in-

[1]Fisher (1925) institutionalized the significance levels. R. A. Fisher supposedly decided to establish the 5% level since he was earning 5% of royalties for his publications. It is however noticeable that, in economics, the academic community has converged toward 10% as being the first hurdle to pass, maybe because of the stringency of the 5% one.

[2]The choice of the right specification may depend on its capacity to detect an effect. For instance, authors may stop exploring further specifications when finding a "significant" one. See, for instance, Bastardi et al. (2011), Nosek et al. (2012).

consistent with an increasing probability of being published, and this inconsistency to be related with researchers' behaviors? Imagine that there are three types of results, green lights are clearly rejected tests, red lights clearly accepted tests and amber lights uncertain tests, i.e. close to the 5% or 10% statistical significance thresholds but not there yet. Assume (i) that it is easier for researchers to produce a green test when first confronted with an amber one rather than with a red one and (ii) that the marginal gains of turning from amber to green are higher than changing from red to amber. In this case, there would be a shift in the observed distribution of statistics due to amber tests being transformed into green tests and this pattern would be inconsistent with our assumption on selection. Indeed, there would be a shortage of amber tests relatively to green tests, which is consistent with selection, but there would also be a shortage of amber tests relatively to red tests. Graphically, we should observe (i) a first bump, (ii) a valley (not enough tests with p-values around 0.15 as if they were disliked relatively to tests with p-values of 0.30) and (iii) the echoing bump (too many tests with p-values slightly under 0.05 as if they were preferred to tests with p-values below 0.001).

We find empirical evidence for this two humped pattern. The distribution of test statistics published in three of the most prestigious economic journals over the period 2005–2011 exhibits a sizable under-representation of marginally insignificant statistics relatively to significant statistics but also to (very) insignificant ones. In a nutshell, once tests are normalized as z-statistics, the distribution has a two humped camel shape with (i) a first hump for low z-statistics, (ii) missing z-statistics between 1.2 and 1.65 (p-values between 0.25 and 0.10) with a local minimum around 1.5 (p-value of 0.12), and (iii) a second hump between 2 and 4 (p-values slightly below 0.05). Our accounting framework allows us to show that this non-monotonic pattern cannot be explained by selection alone under the assumption that selection should be weakly increasing in the z-statistics. There is a large residual that we henceforth refer to as inflation. We find that 10% to 20% percent of tests with p-values between 0.05 and 0.0001 are *misallocated*: there are missing test statistics just before the 0.10 threshold that we can retrieve after the 0.05 threshold.[3]

[3]It is theoretically difficult to separate the estimation of behavioral response of researchers from selection: one may interpret selection and its response as the equilibrium outcome of a game played by editors/referees and authors as in the model of Henry (2009). Editors and referees prefer to publish results that are "significant". Authors are tempted to inflate (with a cost), which pushes editors toward being even more conservative thereby exacerbating selection and inflation. A strong empirical argument in favor of this game between editors/referees and authors would be an increasing selection even

The two-humped camel shape is unlikely to be due to journals favoring green tests and red tests over amber tests. Indeed, we collect a broad range of information on each paper and author and compare the distribution of published tests along various dimensions. While the two-humped shape is an empirical regularity that can be observed consistently across journals, years and fields, it is much less pronounced in articles where stars are not used as eye-catchers. To make a parallel with central banks, the choice not to use eye-catchers might be considered as a commitment from authors to keep inflation low.[4] Importantly, we show that the pattern we document is not driven by the co-existence of null and positive empirical results. Similarly, the two-humped camel shape is less visible in articles with theoretical models, articles using data from randomized control trials or laboratory experiments and papers published by tenured and older researchers. More generally, we find a larger residual in cases in which we would expect higher incentives for researchers to respond to selection.

While many papers compare the density of marginally insignificant tests to marginally significant tests, there is surprisingly little work incorporating the whole distribution of p-values in the analysis. To our knowledge, this project is the first paper that focuses on the two-humped pattern in the distribution of published tests. Using the whole distribution of p-values, we propose an accounting framework to measure what can be explained by selection and what cannot. To achieve this, we collect a very large number of test statistics published in the American Economic Review, the Journal of Political Economy, and the Quarterly Journal of Economics between 2005 and 2011. This collecting process generated $50,078$ tests grouped in $3,389$ tables (or results subsections) and 641 articles. This large number of tests allows us to uncover subtle patterns in the distribution of published tests, and perform subsample analyses by authors and articles characteristics.

The literature on tests in economics was flourishing in the 1980s and has already shown the importance of strategic choices of specifications from authors.[5] For instance, Leamer and Leonard (1983) and Leamer (1985) point out the fact that inferences drawn from coefficients estimated in linear regressions are very sensitive to the underlying econometric model. They

below 0.05, i.e. editors challenge the credibility of rejected tests. Our findings do not seem to support this pattern.

[4]However, such a causal interpretation might be challenged: researchers may give up on stars precisely when their use is less relevant, either because coefficients are very significant and the test of nullity is not a particular concern or because coefficients are not significant.

[5]See Lovell (1983), Denton (1985), De Long and Lang (1992) for a discussion on the implications of individual and collective data mining.

suggest to display the range of inferences generated by a set of models. Leamer (1983) rules out the myth inherited from the physical sciences that econometric inferences are independent of priors: it is possible to exhibit both a positive and a negative effect of capital punishment on crime depending on priors on the acceptable specification. More recently, Gelman and Loken (2014) discuss how data analysis choices can be data-dependent even when the tested hypothesis is motivated directly from theoretical concerns. One contribution of our paper is to document a possible outcome of such strategic choices of specifications.

Our paper also relates to the vast literature on the so-called file drawer problem or selection bias: statistics with low values are censored by journals (see Rosenthal 1979 and Stanley 2005 for reviews, and Hedges 1992, Doucouliagos and Stanley 2011 and for a generalized method to identify reporting bias). A large number of publications quantify the extent to which selection distorts published results (see Ashenfelter and Greenstone 2004 or Begg and Mazumdar 1994). Ashenfelter et al. (1999) propose a meta-analysis of the Mincer equation showing a selection bias in favor of significant and positive returns to education. Card and Krueger (1995) and Doucouliagos et al. (2011) are two other examples of meta-analysis dealing with publication bias. More recently, Havránek (2015) uses a meta-analysis of intertemporal substitution estimates and discusses bias that results from selective reporting practices.

The selection issue has also received a great deal of attention in the medical literature (Berlin et al. 1989, Ioannidis 2005, Ridley et al. 2007) and in psychological science (Bastardi et al. 2011, Fanelli 2010a, Simmons et al. 2011). In addition, Auspurg and Hinz (2011), Gerber and Malhotra (2008b), Gerber and Malhotra (2008a), Gerber et al. (2010) and Masicampo and Lalande (2012) collect distributions of tests in journals in sociology, political science and psychology.[6] We differ from that literature in one important dimension as we are not interested in selection by journals *per se* but in the consequences that it may imply on researchers' behavior. This relates our paper to recent empirical work by Franco et al. (2014) on the stage of research production at which selection occurs. Franco et al. (2014) show that most of the selection occurs before submission: authors do not write up, nor submit, null findings.

Finally, our identification method differs from existing methods. Most articles look for discontinuities or bunching around key significance thresholds (e.g. Gerber and Malhotra 2008b) while we are interested in the whole

[6]See Fanelli (2010b) for a related discussion about the hierarchy of sciences.

distribution of test statistics. A recent paper by Simonsohn et al. (2014) also uses a less local analysis. Simonsohn et al. (2014) use distortions in the distribution of p-values below the .05 threshold in order to detect "p-hacking" (too many p-values just below the .05 threshold). In contrast, our method looks at the distortions above the .05 threshold (too few p-values just above the .05 threshold).

Section 2 details the methodology to construct the dataset and provides some information on test meta-data. Section 3 documents the distribution of tests. Section 4 proposes a method to decompose the observed distribution of published statistics into what can be explained by selection and a residual. Finally, we discuss the results of this method in section 5.

2 Data

In this section, we describe the reporting process of tests published in the American Economic Review, the Journal of Political Economy, and the Quarterly Journal of Economics between 2005 and 2011. We then provide some descriptive statistics.

2.1 Reporting process

One major issue is to select test statistics that represent key hypotheses. Most articles in the file-drawer literature collect all statistics because their identification only relies on discontinuities around significance thresholds. Instead, we want to capture the whole distribution of "central" test statistics. In this regard, we use a subjective, or narrative, approach.

In our narrative approach, we consider, as in Gerber and Malhotra (2008a), that not all coefficients reported in tables should be considered as tests of central hypotheses.[7] We identify variables of interest by looking at the tables and table footnotes and by reading the text where the regressions' results are described. We thus omit explicit control variables. In addition, we do not report explicit placebo tests, i.e. statistical tests that the authors expect to fail. In the rare occurrences in which the status of a test was unclear when reading the paper, we prefer to add a non-relevant test than to censor a relevant one. As we are only interested in tests of

[7]In practice, the large majority of those tests are two-sided tests of regression coefficients and are implicitly discussed in the body of the article (i.e. "coefficients are significant"). 85% of collected test are presented using a regression coefficient and its associated standard error. To simplify the exposition we explain the process as if we only had two-sided tests for regression coefficients but the description applies to our treatment of other tests.

central hypotheses of articles, we also exclude descriptive statistics or group comparisons.[8] A specific rule concerns two-stage procedures. We do not report first-stages, except if the first-stage is described by the authors as a major contribution of the article. We also collect separately tests in extensions or robustness tests. Our narrative approach is better described in the Online Appendix.

Our strategy is different from Gerber and Malhotra (2008a) which only includes articles listing a set of specific hypotheses prior to presenting the tests results. One advantage is that we keep a much larger sample of papers. The obvious defect is that our selection implies some subjective choices.

We report numbers exactly as they are presented in articles, i.e. we never round them up or down. We describe in the next subsection how the data is uniformized across the different articles.

We report some additional information on each test, i.e. the issue of the journal, the starting page of the article and the position of the test in the article, its type (one-sided, two-sided, correlation test, etc.) and the status of the test in the article (main or non-main). We prefer to be conservative and only attribute the status of "non-main" statistics if evidence are clearly presented as "complementary," "additional" or "robustness checks." We also keep track of whether authors present some null empirical results as important contribution. Finally, the number of authors, the research field, JEL codes when available, the presence of a theoretical model, the type of data (laboratory experiment, randomized control trials or other), the use of eye-catchers (stars or other formatting tricks such as bold printing), the number of research assistants and researchers the authors wish to thank, the rate of tenure among authors and data and code availability on the website of the journal are also recorded. We do not report the sample size and the number of variables (regressors) as this information is not always provided by the authors. Exhaustive reporting rules are presented in the Online Appendix.

We also collected information from curricula vitae of all the authors who published in the three journals over the period of interest. We gathered information about academic affiliation at the time of the publication, the position at the main institution (assistant professor, associate professor, etc.), whether the author is or was an editor (or a member of an editorial board) of an economic journal, and the year and the institution where the PhD was earned.

[8] A notable exception to this rule was made for experimental papers where results are sometimes presented as mean comparisons across groups.

2.2 Descriptive statistics

The reporting process described above provides 50,078 tests. Journals do not contribute equally: most of the tests come from the American Economic Review, closely followed by the Quarterly Journal of Economics. The Journal of Political Economy provides a little less than a fifth of the sample. Out of the 50,078 tests extracted from the three journals, around 30,000 are rejected at the 10% significance level, 27,000 at 5%, and 21,000 at 1%.

Table 1 gives the decomposition of tests along several dimensions. The average number of test statistics per article equals 78. It is surprisingly high but it is mainly driven by some articles with a very large number of statistics reported. The median article reports 58 statistics and 5 tables. These figures are reasonable as tests are usually diluted in many different empirical specifications. Most papers test two or three central hypotheses. If, for each of the three hypotheses, there are 20 specifications (4 tables and 5 different specifications per table), we would report 60 statistics, a bit more than our median article. In order to alleviate the issue of potential irrelevant statistics, we further adjust the weight of each test statistics by the number of such statistics in the article such that an article with only one statistic contributes as much as one with 300 statistics.

A rough categorization of article into two large research fields reveals that on fourth are macro-oriented while the remaining are micro-oriented. Most articles report positive empirical findings. Papers that report null and mixed (both positive and null) results represent respectively 2% and 13% of the total number of articles. More than half of the articles use eye-catchers defined as the presence of stars or bold printing in a table, excluding the explicit display of p-values. These starry tests represent more than sixty percent of the total number of tests (the average number of tests is higher in articles using eye-catchers). More than seventy percent of tables from which tests are extracted are considered as main. More than a third of the articles in our sample explicitly rely on a theoretical framework but when they do so, the number of tests provided is not particularly smaller than when they do not. Only a fifth of articles are single-authored.[9]

Tests using data from laboratory experiments or randomized control trials constitute a small part of the overall sample. To be more precise, the AER publishes relatively more experimental articles while the QJE

[9]See Card and DellaVigna (2014, 2013) for recent studies about top journals in economics.

seems to favor randomized controlled trials. The overall contribution of both types is equivalent (with twice as many laboratory experiments than randomized experiments but more tests in the latter than in the former).

3 The distribution of tests

In this section, we describe the raw distribution of tests and propose methods to alleviate the over-representation of round values and the potential overweight attributed to articles with many test statistics. We then derive the distribution of test statistics and comment on it.

The collecting process groups three types of measures: p-values, test statistics when directly reported by the authors, and coefficients and standard errors for the vast majority of tests. In order to obtain a homogeneous sample, we transform p-values into the equivalent z-statistics (a p-value of 0.05 becomes 1.96). For tests reported using coefficients and standard errors, we simply construct the ratio of the two.[10] Recall that the distribution of a t-statistic depends on the degrees of freedom, while that of a z-statistic is standard normal. As we are unable to reconstruct the degrees of freedom for all tests, we treat these ratios as if they were following an asymptotically standard normal distribution under the null hypothesis. Consequently, when the sample size is small, the level of rejection we use is not adequate. For instance, some tests for which we associate a z-statistic of $z = 1.97$ might not be rejected at the 5% significance threshold.

The transformation into z-statistic allows us to observe more easily the fat tail of tests (with small p-values). Figure 1(a) presents the raw distribution. Remark that a very large number of p-values end up below the 0.05 significance threshold (more than 50% of tests are rejected at this significance level).

Two potential issues may be raised with the way authors *report* the value of their tests and the way we *reconstruct* the underlying statistics. First, rational numbers that can be expressed as ratios of small integers get over-represented because of the low precision used by authors. For instance, if the estimate is reported to be 0.020 and the standard error is 0.010, then our reconstructed z-statistic would be exactly 2. Second, more than 100 values are reported in some articles against 4 or 5 in others.

[10] These transformations allow us to obtain direct or reconstructed statistics for all but three types of tests collected: (i) tests reported as a zero p-value, (ii) tests reported as a p-value lower than a threshold (e.g. $p < 0.001$), and (iii) tests reported with a zero standard error. These three cases represent 727 tests, i.e. 1.45% of the total sample.

Which weights are we suppose to give to the former and the latter in the final distribution? This issue might be of particular concern as authors might choose the number of tests they report depending on how close or far they are from the thresholds.[11]

To alleviate the first issue, we randomly redraw a number in the interval of potentially true numbers around each collected value. We achieve this by looking at the number of reported digits. In the example given above, the true estimate should lie in the interval $[0.0195, 0.0205]$ and the true standard error in the interval $[0.0095, 0.0105]$ (with a reported estimate of 0.02 instead of 0.020, the interval would have been $[0.015, 0.025]$). We independently redraw an estimate and a standard error in these intervals using a uniform distribution, and then reconstruct a z-statistics thanks to these two random numbers. This reallocation is mostly aesthetic and has very little impact on the analysis: we reallocate z-statistics very close to their initial level. Consequently, we smooth potential discontinuities in histograms but we do not change the overall shape of the distribution.[12] Note, however, that such reallocation would affect a discontinuity analysis around significance thresholds.

To alleviate the second issue, we construct two different sets of weights, accounting for the number of tests per article and per table in each article. For the first set of weights, we associate to each test the inverse of the number of tests presented in the same article such that each article contributes the same to the distribution. For the second set of weights, we associate the inverse of the number of tests presented in the same table (or result sub-section) multiplied by the inverse of the number of tables in the article such that each article contributes the same to the distribution and tables of a same article have equal weights.

Figure 1(b) presents the de-rounded distribution.[13] The shape is striking. The distribution presents a two-humped camel pattern with a local minimum around $z = 1.5$ (p-value of 0.12) and a local maximum around 2 (p-value under 0.05). The presence of a local maximum around 2 is not very surprising, the existence of a valley before more so. Intuitively, the

[11]For example, one might conjecture that authors report more tests when the p-values are closer to the significance thresholds. Conversely, one may also choose to display a small number of satisfying tests as others tests would fail.

[12]For statistics close to significance levels, we could have taken advantage of the information embedded in the presence of a star. However, this approach could only have been implemented for a much reduced number of observations and only in cases where stars are used.

[13]In what follows, we use the word "de-rounded" to refer to statistics to which we applied the method described above.

"natural" distribution of tests, e.g. Student distributions under the null hypothesis, is likely to have a decreasing pattern over the whole interval. On the other hand, selection could explain a monotonically increasing pattern for the distribution of z-statistics at the beginning of the interval $[0, \infty)$. Both effects put together could explain the presence of a unique local maximum, a local minimum before, less so. Our empirical strategy will consist in capturing this non-monotonicity and quantifying this shift in the distribution of tests.[14]

Figures 1(c) and (d) present the weighted distributions of de-rounded statistics. The camel shape is more pronounced than for the unweighted distributions. A simple explanation is that weighted distributions underweight articles and tables for which a lot of tests are reported. For these articles and tables, our way to report tests might have included tests of non-central hypotheses.

The pattern shown in figures 1(b)–(d) is a very robust empirical regularity: the pattern can be seen in any journal and in any year. In addition, as shown by figures presented in the Online Appendix, this empirical regularity is similar between main tests and robustness checks and we do not gain much insights by analyzing separately "complementary," "additional" or "robustness checks." However, the empirical regularity is much less acute when we explicitly select the articles from which we extract our tests by important author or article characteristics. For example, the camel shape is less pronounced in articles without eye-catchers and articles with a theoretical contribution. Similarly, the two-humped camel shape is less pronounced in papers written by senior researchers regardless of whether seniority is captured by years since PhD, tenure or editorial responsibilities. In contrast, it does not vary by data and codes availability on journals' website. This bird's-eye-view across subsamples indicates that there is some heterogeneity in the extent to which amber tests are under-represented and this heterogeneity can be related to author and article characteristics. We

[14]It is important to have in mind that our identification does not rely on the discontinuity around significance threshold but rather on the non-monotonic pattern around this threshold. In the Online Appendix, we nonetheless test for discontinuities. We find evidence that the total distribution of tests presents a small discontinuity around the 0.10 significance threshold, but not much around the 0.05 or the 0.01 thresholds. This modest effect might be explained by (i) our de-rounding process and by (ii) the dilution of hypotheses tested in journal articles. In the absence of a single test, empirical economists provide many converging arguments under the form of different specifications for a single effect. Besides, an empirical article is often dedicated to the identification of more than one mechanism. As such, the real statistic related to an article is a distribution or a set of arguments and this dilution may smooth potential discrepancies around thresholds. The Online Appendix also presents an analysis using Benford's law to look for manipulation in reported coefficients and standard errors.

come back to the analysis of heterogeneity across sub-samples in section 5 where we apply our accounting framework, but this heterogeneity tends to support the interpretation that such a shape results from author behavior.

4 A method to measure inflation

In this section, we provide a purely accounting framework which aims at separating a residual, *inflation*, from *selection*. We first present a very simple descriptive model of selection in academic publishing and define selection in this context. We then decompose any distribution of published statistics into what could be generated by such selection and a residual. Finally, we discuss stories that may challenge our interpretation of this residual, i.e. mechanisms that could enter the residual without reflecting a behavioral response from researchers. We present the intuition of the accounting decomposition in what follows.

For each possible value of test statistics, consider the proportion of submitted working papers that end up being published. If we assume that selection is monotonically increasing with the value of test statistics, all other things being equal, we should observe that the proportion of submitted papers that end up being published increases with the value of test statistics. Non-monotonic patterns in the distribution of test statistics, like in the case of a two humped shape, cannot be explained by selection alone. The valley and the echoing bump that we have uncovered in the previous section will thus be captured by the residual, *inflation*, after estimating how a monotonic selection process fits the observed distribution.

4.1 Notations

We consider a simple process of selection into journals. We abstract from authors and directly consider the universe of working papers as given.[15] Each economic working paper has a unique hypothesis which is tested with a unique specification. Denote by z the absolute value of the statistics associated to this test[16] and $\varphi(.)$ the density of its distribution over the universe of working papers, the *input*.

A unique journal gives a value $f(z, \varepsilon)$ to each working paper where ε

[15]Note that the selection of potential economic issues into a working paper is not modeled here. One can think alternatively that this is the universe of potential ideas and selection would then include the process from the "choice" of idea to publication.

[16]Alternatively, you could see this unique z as the average test statistics of a given paper.

is a noise entering into the selection process in addition to the value z. ε, the noise, may capture a lot of dimensions: the inclinations of journals for certain articles, the importance of the question, the originality of the methodology, the quality of the paper and, most importantly, the behavioral responses of researchers. Working papers are accepted for publication as long as they pass a certain threshold F, i.e. $f(z, \varepsilon) \geq F$. Suppose without loss of generality that f is strictly increasing in ε, such that a high ε corresponds to articles with higher likelihood to be published, for the same z. Denote by G_z the distribution of ε conditional on the value of z. This distribution will capture variations in the probability to be published that are orthogonal to selection. Indeed, the density of tests observed in journals – the output – can be written as:

$$\psi(z) = \frac{\int_0^\infty \left[\mathbb{1}_{f(z,\varepsilon)\geq F} dG_z(\varepsilon) d\varepsilon\right] \varphi(z)}{\int_0^\infty \int_0^\infty \left[\mathbb{1}_{f(z,\varepsilon)\geq F} dG_z(\varepsilon) d\varepsilon\right] \varphi(z) dz}.$$

The observed density of tests $\psi(z)$ for a given z depends on the share of articles with ε sufficiently high to pass the threshold ($\int_0^\infty \left[\mathbb{1}_{f(z,\varepsilon)\geq F} dG_z(\varepsilon) d\varepsilon\right]$) and on the number of such articles, i.e. input ($\varphi(z)$). As the value of z changes, the minimum noise ε required to pass the threshold changes: it is easier to get in, this is the selection effect and it only reflects properties of the function f. In our framework, the distribution G_z of this ε may also change conditionally on z. Any such changes in the distribution of noise to conditional on z will be the residuals to selection. For instance, a shortage of amber tests would correspond to a local shift in the distribution of noise G_z: there would be too many accepted working papers in the green zone (G_z would tilt toward high ε) compared to the amber zone (G_z would tilt toward low ε).

With these notations, changes in observed $\psi(z)$ that cannot be attributed to selection are, by construction, attributed to G_z, the residual.

4.2 Selection

Our empirical strategy consists in estimating how well selection explains differences between the input distribution and the observed distribution of statistics. To this end, we need to characterize selection, i.e. to define the set of selection functions f that can model the attitude of journals.

Let us assume that we know the input distribution φ. The ratio of the

output density to the input density can be written as:

$$\psi(z)/\varphi(z) = \frac{\int_0^\infty \left[\mathbb{1}_{f(z,\varepsilon)\geq F} dG_z(\varepsilon) d\varepsilon\right]}{\int_0^\infty \int_0^\infty \left[\mathbb{1}_{f(z,\varepsilon)\geq F} dG_z(\varepsilon) d\varepsilon\right] \varphi(z) dz}.$$

This quantity $\psi(z)/\varphi(z)$ can be thought of as the proportion of submitted working papers that end up being published for a given z. In this framework, once normalized by the input, the output is a function of the selection function f. We impose the following condition on selection functions:

Assumption 1 (Journals like stars)**.** *The function f is (weakly) increasing in z.*

For a same noisy component ε, journals prefer higher z. Everything else equal, a 10% test is never strictly preferred to a 9% one.

On the one hand, this assumption can be conservative because it may attribute to selection some patterns due to the behavior of researchers. For instance, researchers may censor themselves. Thus, we think that we capture more than only selection by journals, we capture in effect all behaviors that select monotonically along the value of a test statistics.

On the other hand, the assumption that journals prefer tests rejecting the null may not be viable for high z-statistics. Such results could indicate an empirical misspecification to referees. This effect, if present, should only appear for very large statistics. Another concern is that journals may also appreciate clear acceptance of the null hypothesis, in which case the selection function would be initially decreasing. We discuss in greater details the mechanisms challenging our assumption at the end of this section.

In the following lines, we characterize the distributions of statistics that are associated with selection functions satisfying assumption 1. We show that, when G_z is independent from z, there is a correspondence between the set of such functions and the set of observed distributions ψ such that $\psi(z)/\varphi(z)$ is weakly increasing in z.

First, if we shut down any other channels than selection (the distribution of noise is independent of z), there is an increasing pattern in the selection process, i.e. the proportion of articles selected $\psi(z)/\varphi(z)$ should be weakly increasing in z. Indeed, a higher z-statistic is associated with a lower minimum noise ε required to pass the threshold F. As a result, the pool of papers that have an ε sufficiently high to become eligible for publication increases with z. Hence, for any distribution of noise G, the proportion of submitted working papers that end up being published *always*

weakly increases with the value of z and we cannot explain any decrease in this ratio with selection alone: there are proportionally more working papers that end up being published for higher value of test statistics.

Second and this is the purpose of the lemma below, the reciprocal is also true: any increasing pattern for the ratio output to input $\psi(z)/\varphi(z)$ can be rationalized by selection alone, i.e. with a distribution of noise $\tilde{G}$ invariant in z. Any increasing function of z (in a reasonable interval) for the ratio of densities can be generated by a certain function f verifying assumption 1 maintaining the distribution of noise invariant in z. With an increasing ratio of densities, all can be explained by selection.

Lemma 1 (Duality)**.** *Given a selection function f, any increasing function $r : [0, T_{lim}] \mapsto [0, 1]$ (ratio) can be represented by a cumulative distribution of quality $\varepsilon \sim \tilde{G}$, where $\tilde{G}$ is invariant in z:*

$$\forall t, \quad r(z) = \int_0^\infty \left[\mathbb{1}_{f(z,\varepsilon) \geq F} d\tilde{G}(\varepsilon) d\varepsilon \right]$$

$\tilde{G}$ is uniquely defined on the subsample $\{\varepsilon, \exists z \in [0, \infty), f(z, \varepsilon) = F\}$, i.e. on the values of noise for which some articles may be rejected (with insignificant tests) and some others accepted (with significant tests).

Proof. In the Appendix. □

4.3 Estimation

Following this lemma, our empirical strategy consists in the estimation of the best-fitting non-parametric increasing function $\tilde{r} \in Z = \{r \in C([0, \bar{z}]), s.t. r(z) \geq r(z') \Leftrightarrow z \geq z'\}$ for the ratio $\psi(z)/\varphi(z)$. We find the weakly increasing $\tilde{r}$ that minimizes the weighted distance with the ratio $\psi(z)/\varphi(z)$:

$$\min_{\tilde{r} \in Z} \sum_i \left[\psi(z_i)/\varphi(z_i) - \tilde{r}(z_i) \right]^2 \varphi(z_i),$$

where i is the test's identifier. The term $\tilde{r}$ is what can be explained by selection in the ratio of distribution $\psi(z)/\varphi(z)$, but what is the residual of this estimation?

The following corollary relates the error term of the previous estimation to the number of statistics unexplained by selection, i.e. our residual.

Corollary 1 (Residual)**.** *Following the previous lemma, there exists a cumulative distribution $\tilde{G}$ which represents $\tilde{r}$. $\tilde{G}$ is uniquely defined on*

$\{\varepsilon, \exists z \in [0, T_{lim}], f(z,\varepsilon) = F\}$ *and satisfies:*

$$\forall t, \quad \tilde{r}(z) = \frac{\int_0^\infty \left[\mathbb{1}_{f(z,\varepsilon)\geq F} d\tilde{G}(\varepsilon)d\varepsilon\right]}{\int_0^\infty \int_0^\infty \left[\mathbb{1}_{f(z,\varepsilon)\geq F} dG_z(\varepsilon)d\varepsilon\right] \varphi(z)dz}.$$

The residual of the previous estimation can be written as the difference between $\tilde{G}$ *and the true* G_z*:*

$$u(z) = \frac{\tilde{G}(h(z)) - G_z(h(z))}{\int_0^\infty \int_0^\infty \left[\mathbb{1}_{f(z,\varepsilon)\geq F} dG_z(\varepsilon)d\varepsilon\right] \varphi(z)dz},$$

where h *is defined as* $f(z,\varepsilon) \geq F \Leftrightarrow \varepsilon \geq h(z)$*.*

Proof. In the Appendix. □

The quantity u is the residual implied by the component of the observed ratio that cannot be rationalized by a monotonically increasing selection function f and a distribution of noise $\tilde{G}$ independent of z. Indeed, letting $\tilde{\psi}(z) = (1 - \tilde{G}(h(z)))\varphi(z)$ denote the density of z-statistics associated with $\tilde{G}$, the cumulated residual, i.e. the difference between the observed and explained densities of z-statistics, is:

$$\int_0^z \psi(\tau)d\tau - \int_0^z \tilde{\psi}(\tau)d\tau = \int_0^z u(\tau)\varphi(\tau)d\tau.$$

This corollary allows us to map the cumulated error term of the estimation with a quantity that can be easily interpreted: $\int_0^z \psi(\tau)d\tau - \int_0^z \tilde{\psi}(\tau)d\tau$ is the number of z-statistics within $[0, z]$ that cannot be explained by a selection function satisfying assumption 1. When positive (negative), we interpret this residual as an excess (shortage) of z-statistics relative to the input distribution that cannot be explained by a monotonically increasing selection process alone.

To conclude, we have developed a simple accounting framework that decomposes the ratio of observed densities to input into a monotonically increasing selection component and an unexplained component – a residual. We argue that this unexplained component captures, among other things, the behavioral responses of researchers. A difficulty arises in practice. The previous strategy can be implemented non-parametrically for any given input distribution. Which input distribution should we consider? We turn to this question now.

4.4 Input

In the process that occurs before publication, there are several choices that impact the final distribution of tests: the choice of the research question, the dataset, the decision to submit and the acceptance of referees. For instance, Franco et al. (2014) show that some selection occurs before the submission stage: authors do not submit null results. We think that most of these processes are very likely to satisfy assumption 1 (at least for z-statistics that are not extremely high) and these choices, i.e., research question, data analysis choices (see Gelman and Loken (2014)), submission and acceptance (see Franco et al. (2014)), will be captured by the selection process f.

Since we do not observe a "natural" distribution of tests before all choices are made, we consider a large range of distributions for the input. The classes of distribution can be limited to (i) unimodal distributions – with the mode being 0 – because the output for some of our subsamples are unimodal[17] in 0, and (ii) ratio distributions because the vast majority of our tests are ratio tests. They should also capture as much as possible of the fat tail of the observed distribution (distributions should allow for a large number of rejected tests and very high z-statistics). In line with these observations, we consider three candidate classes.

Class 1 (Gaussian)**.** *The Gaussian/Student distribution class arises as the distribution class under the null hypothesis of t-tests. Under the hypothesis that tests are t-tests for independent random processes following normal distributions centered in* 0*, the underlying distribution is a standard normal distribution (if all tests are carried out with infinite degrees of freedom), or a mix of Student distributions (in the case with finite degrees of freedom).*

This class of distributions naturally arises under the assumption that the underlying null hypotheses are always true. For instance, tests of correlations between variables that are randomly chosen from a pool of uncorrelated processes would follow such distributions. However, we know from the descriptive statistics that selection should be quite drastic when we consider a normal distribution for the exogenous input. The output displays more than 50% of rejected tests against 5% for the normal distribution. A normal distribution would rule out the existence of statistics around 10. In order to account for the fat tail observed in the data, we extend the class of exogenous inputs to Cauchy distributions. Remark that the ratio of two normal distributions follows a Cauchy distribution. In that respect, the

[17] If the selection function is increasing and the output is an unimodal distribution in 0, then the input needs to be a unimodal distribution in 0.

class of Cauchy distributions satisfies all the ad hoc criteria that we wish to impose on the input.

Class 2 (Cauchy)**.** *The Cauchy distributions are fat-tail ratio distributions which extend the Gaussian/Student distributions: (i) the standard Cauchy distribution coincides with the Student distribution with* 1 *degree of freedom, (ii) this distribution class is, in addition, a strictly stable distribution.*

Cauchy distributions account for the fact that researchers identify mechanisms among a set of correlated processes, for which the null hypothesis might be false. As such, the Cauchy distribution allows us to extend the input to fat-tail distributions.

The last class are distributions that we derive empirically by performing random tests on large datasets.[18]

Class 3 (Empirical)**.** *We randomly draw variables from a dataset, run* $2,000,000$ *regressions between these variables, and collect the z-statistic behind the first explanatory variable. We apply this procedure to four different datasets: the World Development Indicators (WDI), the Quality of Government dataset (QOG), the Panel Study of Income Dynamics (PSID), and the Vietnam Household Living Standards Survey (VHLSS).*[19]

How do these different classes of distributions compare to the observed distribution of published tests?

Figures 2(a) and (b) show how poorly the normal distribution fits the observed distribution. The assumption that the input comes from uncorrelated processes can only be reconciled with the observed output through a drastic selection (which would generate the observed fat tail from a Gaussian tail). The fit is slightly better for the Student distribution of degree 1. The proportion of rejected tests is then much higher with 44% of rejected tests at the 0.05 significance level and 35% at 0.01. Figures 2(c)–(f) show that the Cauchy distributions as well as the empirical inputs may help to capture the fat tail of the observed distribution. More than the levels of the densities, it is their shape which advocates in favor of the use of these distributions as inputs: if we suppose that selection and inflation become much less intense, if not absent, once we pass a certain threshold, we should indeed observe a constant ratio output/input for these very high z-statistics.

[18] Another potential empirical input could also be the statistics of non-central tests in published papers, such as the significance of control variables for instance.

[19] So, we just ran eight million regressions (see Sala-i Martin 1997 and Hendry and Krolzig 2004).

From that discussion, we extract candidates that cover the range of possible distributions. We keep (i) empirical inputs, (ii) the Student(1) distribution; (iii) and a rather thin-tail distribution, i.e. the Cauchy distribution of parameter 0.5. These distributions cover a large spectrum of shapes and our results are not sensitive to the choice of inputs.

4.5 Discussion

The quantity that we isolate is a cumulated residual (the difference between the observed and the predicted cumulative function of z-statistics) that cannot be explained by a monotonically increasing selection process alone. In our interpretation, it should capture the observed *local* shift of z-statistics that turns amber tests into green ones. This quantity may be a lower bound of inflation as any *globally* increasing pattern (in z) in the inflation mechanism would be captured as part of the selection effect. Only the fact that this inflation is particularly acute for just-insignificant tests is captured here.

Some observations may challenge our interpretation.

First, the selection function by journals may not be increasing because a well-estimated zero, i.e. a null results, might be valued by journals. Our two-humped shape may come from the aggregation of two very different types of papers, one with a mode around 0, and one with a mode around 2. Indeed, some editors may favor well-estimated zeros or well-estimated zeros could be valued differently across fields. The first hump would then come from papers associated to one type of editors or one group of research fields and our second hump would be associated with the other type. In order to discard this interpretation, we perform a series of subsample decompositions. In a first step, we isolate the 16 papers having stated a null result as their main contribution. As shown by figure 3(a), the distribution of z-statistics for papers with a null result for the main hypothesis is unimodal with a mode around 0. In spite of the markedly different pattern shown in these papers, there are not enough of them to explain the presence of a first hump in the whole distribution. Indeed, in figure 3(b), we exclude all papers in which there is *at least one null result* presented for one important hypothesis (15% of the sample), and the shape is remarkably similar to our benchmark distribution. We also perform a decomposition by economic fields to check that our results are not driven by the presence of well-estimated zeros in a particular field. One may think that applied microeconomists put more emphasis on the precision of estimates

(large number of observations) and the method (random or quasi-random experiments) while there are usually less observations in macro analyses. Figures 3(c) and (d) display the decomposition between micro- and macro-oriented articles. We find similar patterns for the two distributions. Further (unreported) decompositions into more disaggregated subfields, e.g., labor, development or trade, show no singularity in any subfield. Finally, the preference for well-estimated zeros should not depend on article and author characteristics. Yet, as shown later, we find that distributions of z-statistics vary with these features and we can find subsamples in which the second hump is almost absent.

Second, imagine that the authors could predict where their tests will end up and decide to invest in the empirical investigation accordingly. This *ex ante* selection is captured by the selection term as long as it displays an increasing pattern, i.e. projects with expected higher z-statistics are more likely to be undertaken. There is a very common setting in which it is unlikely to be the case: when designing experiments (or randomized control trials), researchers compute power tests such as to derive the minimum number of participants for which an effect can be statistically captured. Experiments are expensive and costs need to be minimized under the condition that a test may settle whether the hypothesis can or cannot be rejected. We should expect a thinner tail for those experimental settings and this is exactly what we observe. In such instance, our assumptions fail and the results that we produce are to be taken with a grain of salt because the residual then captures these missing large z-statistics. We think that, for "reasonable" z-statistics, such behavior is unlikely in non-experimental cases because of the limited capacity of authors to predict where the z-statistics may end up as well as the modest incentives to limit oneself to small samples, and indeed we find this "thinner tail" pattern only in the experiments-RCT subsample. However, in order to limit the influence of very large z-statistics, we will restrict our analysis to z-statistics below 10.

Third, our main assumption that any test with a weakly higher z-statistic has a higher likelihood of being accepted by a journal requires that we condition for other heterogeneity across articles. Indeed, when some papers test a novel relationship, the standards of acceptance may be lower. We cannot control directly for such unobservable heterogeneity. Instead, we will show how our results differ between sub-samples chosen along some important observable characteristics, e.g. being tenured and age.

5 Results

In this section, we first apply our estimation strategy to the full sample and propose non-parametric and parametric analyses. Then, we divide tests into sub-samples and we provide the results separately for each sub-sample.

5.1 Non-parametric application

We group observed z-statistics by bandwidth of 0.01 and limit our study to the interval $[0, 10]$. Accordingly, the analysis is made on $1,000$ bins. We estimate the best increasing fit of the ratio of densities thanks to the Pool-Adjacent-Violators Algorithm.

Figures 4(a)–(f) plot the best increasing fit for the ratio of observed density to the density of the different inputs, and the associated cumulative residual.[20]

Two interpretations emerge from these estimations. First, the best increasing fit $\tilde{f}$ displays high marginal returns to the value of statistics $\partial\tilde{f}(z)/\partial z$ only for $z \in [1.5, 2]$. The marginal returns are 0 otherwise. Selection is intense precisely where it is supposed to be discriminatory, i.e. just before (or between) the thresholds. Second, the misallocation of z-statistics captured by the cumulated residuals starts to increase slightly before $z = 2$ up to 4. In other words, the bulk between p-values of 0.05 and 0.0001 cannot be explained by an increasing selection process alone. At the maximum, the misallocation reaches 0.028 when using the WDI input, which means that 2.8% of the total number of statistics are misallocated between 0 and 4. As there is no residual between 0 and 2, we compare this 2.8% to the total proportion of z-statistics between 2 and 4, i.e. 30% of the total population. The conditional probability of being misallocated for a z-statistic between 2 and 4 is thus around 9%. As shown by figures 4(a)–(f), results do not change depending on the chosen input distribution. Results are very similar both in terms of shape and magnitude. The upper part of table 2 summarizes the results of the estimations by providing the maximum cumulated residual. We can note that our estimates are remarkably consistent across the different input distributions.

[20]Note that there are less and less z-statistics per bins of width 0.01. On the right-hand part of the figures, we can see lines that look like raindrops on a windshield. Those lines are bins for which there is the same number of observed z-statistics. As this observed number of z-statistics is divided by a decreasing and continuous function, this gives these increasing patterns.

A concern with this estimation strategy is that the misallocation could reflect different levels of quality between articles with z-statistics between 2 and 4 compared to the rest. We cannot rule out this possibility. However, two observations gives support to our interpretation: the start of the misallocation is right after (i) the first significance threshold, and (ii) the zone where the marginal returns of the selection function are the highest.[21]

As already suggested by the shapes of weighted distributions, the results are stronger when the distribution of observed z-statistics is corrected such that each article or each table contributes the same to the overall distribution.

Table 2 also presents maximum cumulated residuals obtained when using raw data, i.e. statistics that have not been de-rounded, and data smoothed using exclusion of low-precision values as an alternative smoothing method.[22] The associated results are very close to the previous ones, which illustrates the fact that de-rounding is mostly aesthetic.

Even though our results are strongly inconsistent with the presence of only selection, the distribution of misallocated z-statistics is a little surprising (and not completely consistent with inflation): the surplus observed between 2 and 4 is here compensated by a deficit after 4. Inflation would predict such a deficit before 2 (between 1.2 and 1.7, which corresponds to the valley between the two bumps). This result comes from the fact that inflation is not well non-parametrically identified. We impose that any weakly increasing pattern observed in the ratio of densities should be attributed to the selection function. For instance, the stagnation of the ratio observed before 1.7 is captured by the selection function while one may think that such stagnation is due to inflation. Nonetheless, as the missing tests still fall in the bulk between 2 and 4, they allow us to identify a violation of the presence of selection alone: the bump is too big to be reconciled with the tail of the distribution. In the next sub-section, we eliminate this inconsistency by imposing parametric restrictions on the selection process.

[21] This result is not surprising as it comes from the mere observation that the observed ratio of densities reaches a maximum between 2 and 4.

[22] We define low precision values as values reported with a precision equal to 1, where the precision of a reported number is the number of digits that follows the first non-zero digit. In the case of statistics reported as the ratio of an estimate to a standard error, we define the precision of values as the minimum precision of the two parts of the ratio. For example, 0.001 has precision 1, whereas 2.03 has precision 3. We exclude $11,302$ observations, i.e. about 22% of the sample, according to this criterion.

5.2 Parametric application

A concern about the previous analysis is that the surplus of misallocated tests between 2 and 4 is implicitly compensated by missing tests after this bulk, namely a shortage of tests rejected at very high level of significance. The mere observation of the distribution of tests does not give the same intuition. Apart from the bulk between 2 and 4, the other anomaly is the valley around $z = 1.5$. This valley is considered as a stagnation of the selection function in the previous non-parametric case. We consider here a less conservative test by estimating the selection function under the assumption that it should belong to a set of parametric functions.

Assume now that the selection process can be approximated by an exponential polynomial function, i.e. consider a selection function of the following form:

$$f(z) = c + \exp(a_0 + a_1 z + a_2 z^2).$$

The pattern of this function allows us to account for the concave pattern of the observed ratio of densities.[23]

Figure 5(a)–(f) presents the best parametric fits and the cumulative sums of residuals when using the different inputs. Contrary to the non-parametric case, the misallocation of statistics starts after $z = 1$ (p-values around 0.30) and is decreasing up to $z = 1.65$ (p-values equals to 0.10 and first significance threshold). These missing statistics are then completely retrieved between 1.65 and 4, and no misallocation is left for the tail of the distribution. This statement holds true for all inputs, but there is an additional misallocation before $z = 1$ for Student and Cauchy inputs. This is due to the very large number of small z-statistics generated by these two distributions that cannot be perfectly replicated under our assumption that selection is increasing. In general, however, and as shown in the bottom part of table 2, the magnitude of misallocation is very similar to the non-parametric case.

Overall, the pattern of the cumulated residuals observed in these figures is very consistent with our story: once we account for selection, we identify a shortage of marginally insignificant tests that is compensated with an excess of marginally significant results.

[23]The analysis can be made with simple polynomial functions but it slightly worsens the fit.

5.3 Sub-sample analysis

The information we collected about articles and authors allow us to split the full sample of tests into sub-samples along various dimensions and to compare our measure of inflation across sub-samples. It seems reasonable to expect inflation to vary along characteristics of the paper, e.g. the importance of the empirical contribution, or characteristics of the authors, e.g. the expected returns from a publication in a prestigious journal.[24]

In this sub-section, we split the full sample of published z-statistics along various dimensions and perform a different estimation of the best-fitting selection function on each sub-sample using the methods presented above. For space consideration, we restrict ourselves to the analysis of de-rounded unweighted distributions using the WDI input. Table 3 presents the maximum cumulated residuals from these estimations.[25]

We start with the formal analysis of distributions of statistics when splitting the sample into two rough categories: microeconomics and macroeconomics. Maximum cumulated residuals are remarkably similar in both fields. This suggests that the behavior we are documenting is not due to heterogeneity across fields. The formal analyses of further (unreported) decompositions into finer research fields lead to the same conclusion. We continue by isolating articles that do not present *any* null result as a contribution. We contrast this analysis with the one of articles whose authors put forward a null empirical result as their main contribution. While the number of such articles is way to small to satisfy the requirements of our accounting method, we find that isolating papers with positive results only does not change the results. This finding rules away that the our first hump is due to the presence of null results presented as a main contribution.[26]

In sub-samples presented in figures 6(a) and (b), we split the full sample depending on the presentation of the results and the content of the paper. When distinguishing between tests presented using eye-catchers or not, the analysis shows that the conditional probability of being misallocated for

[24]This analysis cannot be considered as causal. From the blank page to the published research article, researchers choose the topic, collect data, decide on co-authorship, where to submit the paper, etc. All these choices are made either simultaneously or sequentially. None of them can be considered as exogenous since they are related to the expected quality of the outcome and to its expected likelihood to be accepted for publication.

[25]The distribution of statistics within each sub-sample, as well as the associated non-parametric and parametric estimations are presented in the Online Appendix where supplementary tables also present the maximum cumulated residuals for the same sub-samples but using other inputs.

[26]The (unreported) distribution of statistics from paper that explicitly put forward mixed results naturally exhibits a strong camel shape.

a z-statistic between 2 and 4 is around 12% in the eye-catchers sample against 5% in the no eye-catchers sample. Not using stars may act as a commitment for researchers not to be influenced by the distance of their tests from the 10% or 5% significance thresholds. Then, we split the sample depending on whether the test is presented as a main test or not (tests or results explicitly presented as "complementary", "additional" or "robustness checks"). The maximum cumulated residual is around twice as large for results not presented as a main result. The emphasis put on the empirical analysis may also depend on the presence of a theoretical contribution. In articles having a theoretical content, the main contribution of the paper is divided between theory and empirics and the estimation may be constrained by the model. These intuitions may explain the shapes of figures 6(c) and (d): inflation is quite low in articles with a theoretical model compared to articles that do not offer an explicit theoretical contribution.

One might consider that articles and ideas from researchers with higher academic rankings are more likely to be valued by editors and referees. Accordingly, inflation may vary with authors' status: well-established researchers facing less intense selection should have less incentives to inflate. A first proxy that we use to capture authors' status is experience. We compare articles having an average PhD-age of authors below and above the median PhD-age of the sample. We find that inflation is more pronounced among relatively younger authors. A second indicator reflecting authors' status is whether they are involved in the academic editorial process. Accordingly, we split the sample in two groups: the first is made of articles published by authors who were not editors or members of editorial boards before publication, while the second is made of articles published by at least one editor or member of an editorial board. Inflation appears to be slightly larger in the first group. Another proxy of authors' status which is strongly related to incentives to publish in top journals is whether authors are tenured or not. We compute the rate of tenure among authors of each article and split the sample along this dimension.[27] We find that the presence of at least one tenured researcher among authors is associated with a strong decline in inflation. All in all, these findings seem in line with the idea that inflation is likely to vary along expected returns to publication in prestigious journals.

[27]Getting information about effective tenure status of authors may be difficult as position denomination varies across countries and institutions. Here, we only consider full professors as tenured researchers. Furthermore, the length of the publication process makes it hard to know the precise status of authors at the time of submission. Here, we arbitrarily consider positions of authors three years before publication.

We continue by splitting the sample of published tests between single-authored and co-authored papers: inflation is larger in single-authored papers. We collected the number of individuals the authors thank and the number of research assistants mentioned in the published version of the paper: inflation seems to be smaller when no research assistants are acknowledged and in articles with a relatively low number of thanks.

Whether data and codes are available on the website of the journal for replication purposes has attracted a great deal of attention lately (see Dewald et al. 1986 and McCullough et al. 2008). For instance, the AER implemented a mandatory data and code archive few years ago. On the other hand, the JPE archive access is available solely to JPE subscribers. We check for each article whether data and codes are available on the website of the journal. The analysis of the different sub-samples does not show conclusive evidence that data or programs availability mitigate inflation.

To conclude this sub-sample analysis, we investigate the distribution of tests depending on the source of data. There is an increasing use of randomized control trials and laboratory experiments in economics and many researchers advocate this is a very useful way to accumulate knowledge without relying on questionable specifications. As argued earlier, our methodology is not adapted to these sources of data: experiments are designed such as to minimize costs while being able to detect an effect and, by construction, large z-statistics are not likely to appear. Our residual essentially captures this absence of large z-statistics. Indeed, we find that inflation is large, which contrasts with the fact that the distribution of z-statistics for randomized control trials and laboratory experiments does not exhibit a two humped shape as shown by figure 6(e). A visual investigation of the distribution reveals that there is neither a valley between 0.25 and 0.10, nor a significant bump around 0.05, but the tail is much thinner and there are almost no z-statistics after 5. These results confirm the findings by Vivalt (2015) that specification searching and publication biases are quite small in randomized controlled trials.

Overall, we find that the intensity of inflation varies along different dimensions of paper and author characteristics. Interestingly, these variations seem consistent with the returns to displaying a "significant" empirical analysis : these evidence give more support to our interpretation in terms of behavioral responses of researchers.

6 Conclusion

He who is fixed to a star does not change his mind. (Da Vinci)

There exists substantial information asymmetry between the authors of an article and the rest of the academic community. Consequently, Olken (2015) writes that it is believed that "[researchers] are inherently biased and data mine as much as possible until they find results." This belief has direct implications on the behavior of referees and editors who tend to ask for example for the hidden specifications as robustness checks.

In this paper, we have identified a misallocation in the distribution of the test statistics in some of the most respected academic journals in economics. Our analysis suggests that the pattern of this misallocation is consistent with what we dubbed an inflation bias: researchers might be tempted to inflate the value of those almost-rejected tests by choosing a slightly more "significant" specification. We have also quantified this inflation bias: among the tests that are marginally significant, 10% to 20% are misreported. These figures are likely to be lower bounds of the true misallocation as we use conservative collecting and estimating processes. On the one hand, our results provide some evidence consistent with the existence of p-hacking thereby justifying the increasing concerns on data replicability and the implementation of pre-analysis plans. In particular, we have identified paper and author characteristics that seem to be related to the inflation bias, e.g., the use of eye-catchers or being in a tenure-track job. The inflation bias is also associated with the type of empirical analysis (e.g. randomized control trials) and the existence of a theoretical contribution. On the other hand, while our missing tests represent a non-negligible share of the whole population of tests, the bias remains circumscribed (z-statistics from 1.4 to 2.2).

The external validity of our findings is unclear. Our analysis is restricted to three top economic journals. In these journals, the rejection rates are high and the returns to publication are much higher than in other journals. Some researchers with negative results may send their papers to less prestigious journals, and the distribution of tests in the universe of journals may be less biased than in our distribution. Negative results would then benefit from less impact but would still contribute to the literature. Moreover, as opposed to pharmaceutical trials, incentives for data mining are essentially private in economics (career concerns), and our findings may not translate to other disciplines (Olken 2015).

As noted by Fanelli (2009) who discusses explicit professional misconducts, concerns about the existence of an inflation bias are shared in other sciences. These concerns naturally gave birth to calls to reduce the selection and inflation biases and tilt the balance towards "getting it right" rather than "getting it published" (see Weiss and Wagner 2011 and Nosek et al. 2012 among others). For instance, journals (the Journal of Negative Results in BioMedecine or the Journal of Errology) have been launched with the ambition of giving a place where authors may publish non-significant findings. Alternative solutions may rely on sealed-envelope submissions (Dufwenberg and Martinsson 2014). Similarly, pre-analysis plans have been proposed (and used) in natural sciences, but also in social sciences (see Miguel et al. 2014 and Olken 2015 for economics), to reduce data mining. In this paper, we provide evidence that academic economists respond to publication incentives, which justifies these concerns. While the distortion we document can been considered as moderate (Olken 2015), it would be very interesting to replicate our methodology to other disciplines where the incentives are thought to be more distorted, e.g., in medicine with the FDA approval processes. Furthermore, it seems important to investigate whether and how researchers' behavior changed following the implementation of the above mentioned policies. This would echo the distant call by Mahoney (1977) who pointed out that understanding the effects of norms requires not only the identification of the biases, but also an understanding of how the academic community adapts its behavior to those norms, and how beliefs evolve with such adaptation.

References

Ashenfelter, O. and Greenstone, M.: 2004, Estimating the value of a statistical life: The importance of omitted variables and publication bias, *American Economic Review* **94**(2), 454–460.

Ashenfelter, O., Harmon, C. and Oosterbeek, H.: 1999, A review of estimates of the schooling/earnings relationship, with tests for publication bias, *Labour Economics* **6**(4), 453–470.

Auspurg, K. and Hinz, T.: 2011, What fuels publication bias? theoretical and empirical analyses of risk factors using the caliper test, *Journal of Economics and Statistic* **231**(5-6), 636 – 660.

Bastardi, A., Uhlmann, E. L. and Ross, L.: 2011, Wishful thinking, *Psychological Science* **22**(6), 731–732.

Begg, C. B. and Mazumdar, M.: 1994, Operating characteristics of a rank correlation test for publication bias, *Biometrics* **50**(4), pp. 1088–1101.

Berlin, J. A., Begg, C. B. and Louis, T. A.: 1989, An assessment of publication bias using a sample of published clinical trials, *Journal of the American Statistical Association* **84**(406), 381–392.

Card, D. and DellaVigna, S.: 2013, Nine facts about top journals in economics, *Journal of Economic Literature* **51**(1), 144–61.

Card, D. and DellaVigna, S.: 2014, Page limits on economics articles: Evidence from two journals, *Journal of Economic Perspectives* **28**(3), 149–68.

Card, D. and Krueger, A. B.: 1995, Time-series minimum-wage studies: A meta-analysis, *The American Economic Review* **85**(2), 238–243.

De Long, J. B. and Lang, K.: 1992, Are all economic hypotheses false?, *Journal of Political Economy* **100**, 1257–1257.

Denton, F. T.: 1985, Data mining as an industry, *The Review of Economics and Statistics* **67**(1), 124–27.

Dewald, W. G., Thursby, J. G. and Anderson, R. G.: 1986, Replication in empirical economics: The journal of money, credit and banking project, *American Economic Review* **76**(4), 587–603.

Doucouliagos, C. and Stanley, T. D.: 2011, Are all economic facts greatly exaggerated? theory competition and selectivity, *Journal of Economic Surveys* **27**(2), 316–339.

Doucouliagos, C., Stanley, T. and Giles, M.: 2011, Are estimates of the value of a statistical life exaggerated?, *Journal of Health Economics* **31**(1).

Dufwenberg, M. and Martinsson, P.: 2014, Keeping researchers honest: The case for sealed-envelope-submissions, *Technical Report 533*, IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University.

Fanelli, D.: 2009, How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data, *PLoS ONE* **4**(5).

Fanelli, D.: 2010a, Do pressures to publish increase scientists' bias? an empirical support from us states data, *PLoS ONE* **5**(4).

Fanelli, D.: 2010b, "positive" results increase down the hierarchy of the sciences, *PLoS ONE* **5**(4).

Fisher, R. A.: 1925, *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh.

Franco, A., Malhotra, N. and Simonovits, G.: 2014, Publication bias in the social sciences: Unlocking the file drawer, *Science* **345**(6203), 1502–1505.

Gelman, A. and Loken, E.: 2014, The statistical crisis in science, *American Scientist* **102**, 460.

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

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

Gerber, A. S., Malhotra, N., Dowling, C. M. and Doherty, D.: 2010, Publication bias in two political behavior literatures, *American Politics Research* **38**(4), 591–613.

Havránek, T.: 2015, Measuring intertemporal substitution: The importance of method choices and selective reporting, *Journal of the European Economic Association* **Forthcoming**.

Hedges, L. V.: 1992, Modeling publication selection effects in meta-analysis, *Statistical Science* **7**(2), 246–255.

Hendry, D. F. and Krolzig, H.-M.: 2004, We ran one regression, *Oxford Bulletin of Economics and Statistics* **66**(5), 799–810.

Henry, E.: 2009, Strategic disclosure of research results: The cost of proving your honesty, *Economic Journal* **119**(539), 1036–1064.

Ioannidis, J. P. A.: 2005, Why most published research findings are false, *PLoS Med* **2**(8), e124.

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

Leamer, E. E.: 1985, Sensitivity analyses would help, *The American Economic Review* **75**(3), 308–313.

Leamer, E. and Leonard, H.: 1983, Reporting the fragility of regression estimates, *The Review of Economics and Statistics* **65**(2), pp. 306–317.

Lovell, M. C.: 1983, Data mining, *The Review of Economics and Statistics* **65**(1), 1–12.

Mahoney, M. J.: 1977, Publication prejudices: An experimental study of confirmatory bias in the peer review system, *Cognitive Therapy and Research* **1**(2), 161–175.

Masicampo, E. J. and Lalande, D. R.: 2012, A peculiar prevalence of p values just below .05, *Quarterly Journal of Experimental Psychology* pp. 1–9.

McCullough, B., McGeary, K. A. and Harrison, T. D.: 2008, Do economics journal archives promote replicable research?, *Canadian Journal of Economics* **41**(4), 1406–1420.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K., Gerber, A., Glennerster, R., Green, D., Humphreys, M., Imbens, G. et al.: 2014, Promoting transparency in social science research, *Science* **343**(6166), 30–31.

Nosek, B. A., Spies, J. and Motyl, M.: 2012, Scientific utopia: Ii - restructuring incentives and practices to promote truth over publishability, *Perspectives on Psychological Science* **7**(6), 615–631.

Olken, B.: 2015, Pre-analysis plans in economics. Unpublished manuscript.

Ridley, J., Kolm, N., Freckelton, R. P. and Gage, M. J. G.: 2007, An unexpected influence of widely used significance thresholds on the distribution of reported p-values, *Journal of Evolutionary Biology* **20**(3), 1082–1089.

Rosenthal, R.: 1979, The file drawer problem and tolerance for null results, *Psychological Bulletin* **86**, 638.

Sala-i Martin, X.: 1997, I just ran two million regressions, *American Economic Review* **87**(2), 178–83.

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

Simonsohn, U., Nelson, L. D. and Simmons, J. P.: 2014, P-curve: A key to the file drawer, *Journal of Experimental Psychology: General* **143**, 534–547.

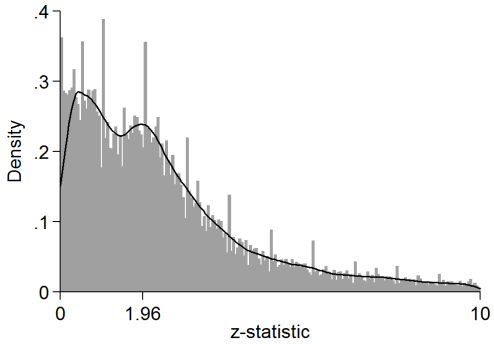
Stanley, T. D.: 2005, Beyond publication bias, *Journal of Economic Surveys* **19**(3), 309–345.

Sterling, T. D.: 1959, Publication decision and the possible effects on inferences drawn from tests of significance-or vice versa, *Journal of The American Statistical Association* **54**, pp. 30–34.

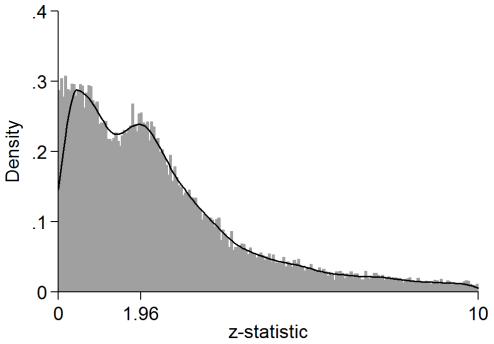
Vivalt, E.: 2015, How much can we generalize from impact evaluations? Unpublished manuscript.

Weiss, B. and Wagner, M.: 2011, The identification and prevention of publication bias in the social sciences and economics, *Journal of Economics and Statistics* **231**(5 - 6), 661 – 684.

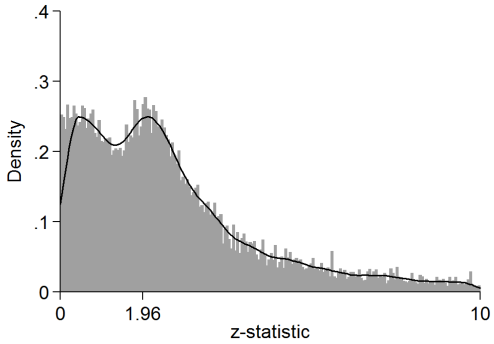
Figure 1: Distributions of z-statistics.



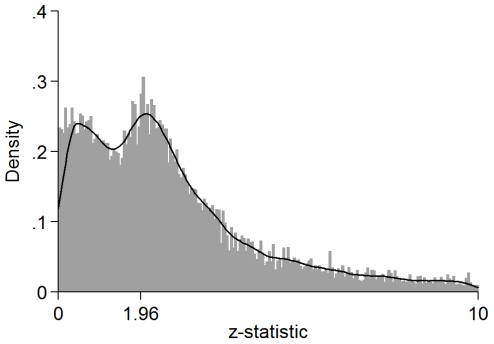
(a) Raw distribution of z-statistics.



(b) De-rounded distribution of z-statistics.



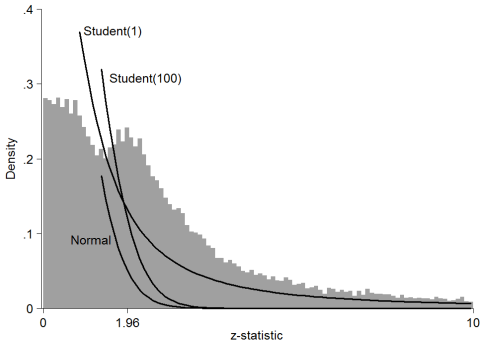
(c) De-rounded distribution of z-statistics, weighted by articles.



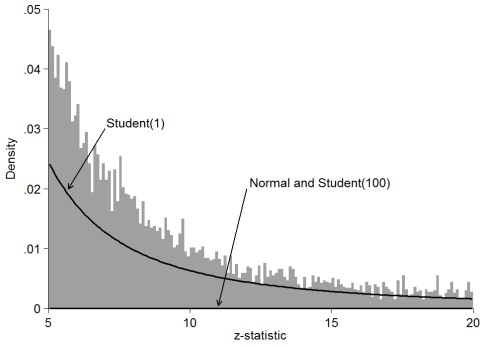
(d) De-rounded distribution of z-statistics, weighted by articles and tables.

Sources: AER, JPE, and QJE (2005-2011). See the text for the de-rounding method. The distribution presented in sub-figure (c) uses the inverse of the number of tests presented in the same article to weight observations. The distribution presented in sub-figure (d) uses the inverse of the number of tests presented in the same table (or result) multiplied by the inverse of the number of tables in the article to weight observations. Lines correspond to kernel density estimates.

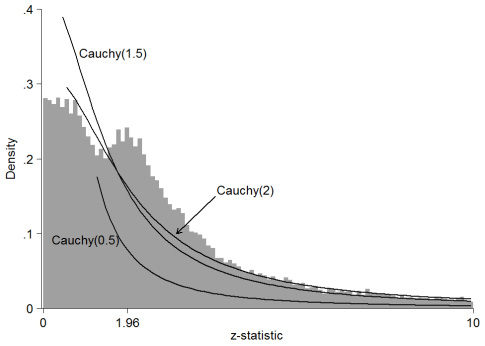
Figure 2: Distribution of z-statistics and candidate exogenous inputs.



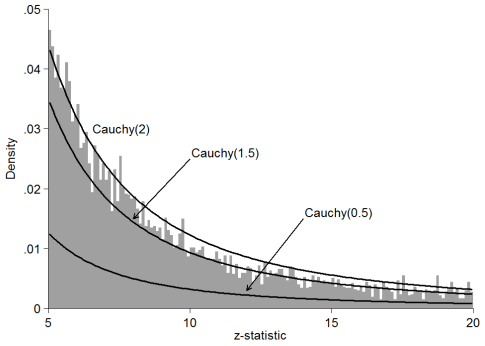
(a) Gaussian/Student inputs ($0 < z < 10$).



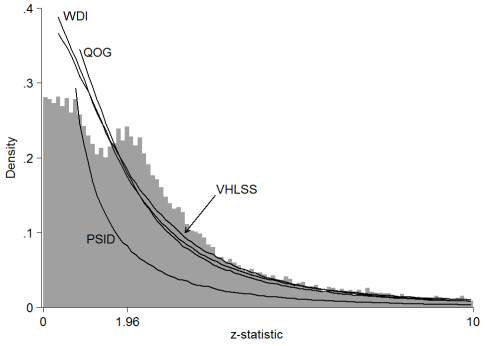
(b) Gaussian/Student inputs ($5 < z < 20$).



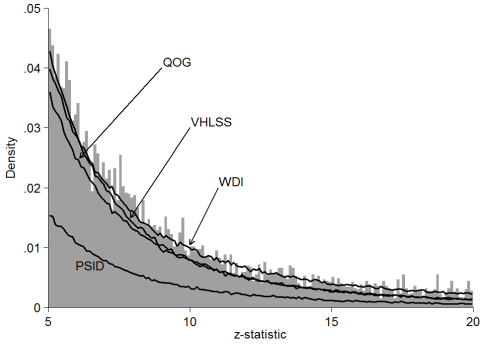
(c) Cauchy inputs ($0 < z < 10$).



(d) Cauchy inputs ($5 < z < 20$).



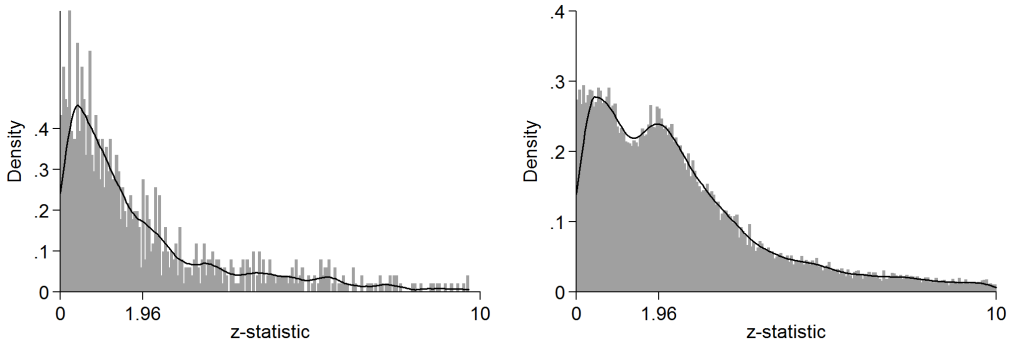
(e) Empirical inputs ($0 < z < 10$).



(f) Empirical inputs ($5 < z < 20$).

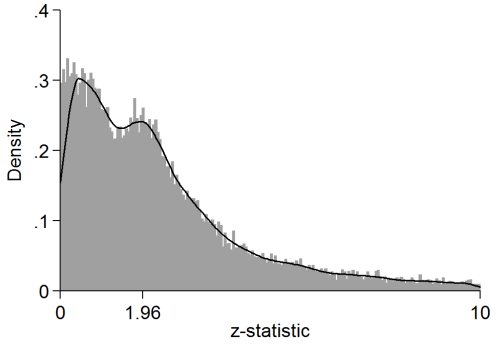
Sources: AER, JPE, and QJE (2005-2011). Unweighted distributions plotted using de-rounded statistics.

Figure 3: Distributions of z-statistics for different sub-samples: microeconomics *versus* macroeconomics, and nature of empirical evidence.

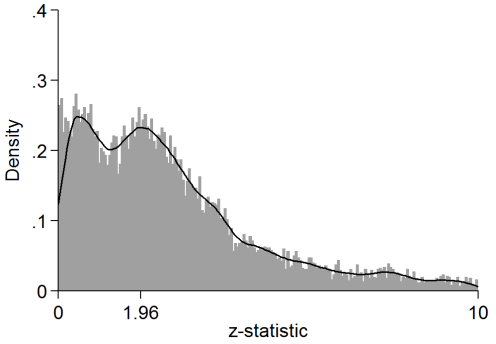


(a) Null results.

(b) Positive results.



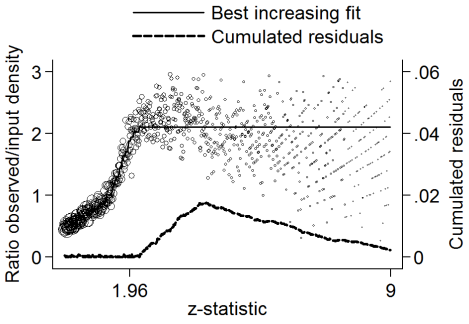
(c) Microeconomics.



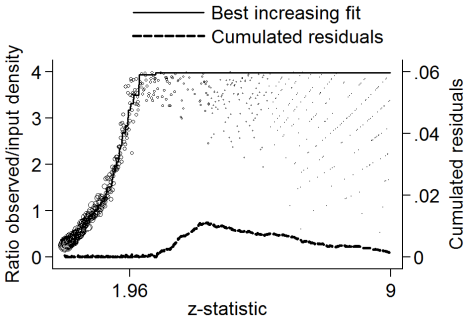
(d) Macroeconomics.

Sources: AER, JPE, and QJE (2005-2011). Distributions are unweighted and plotted using de-rounded statistics. Lines correspond to kernel density estimates.

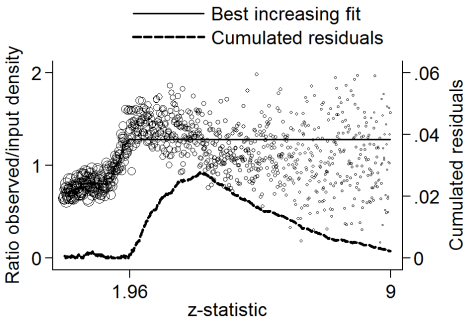
Figure 4: Non-parametric estimation of selection and inflation.



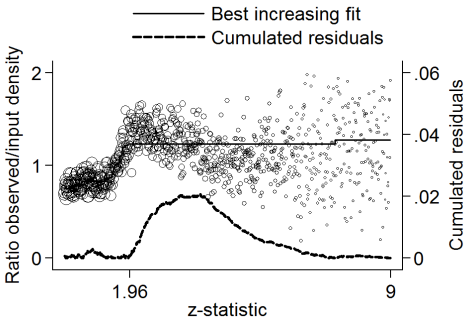
(a) Student(1) input.



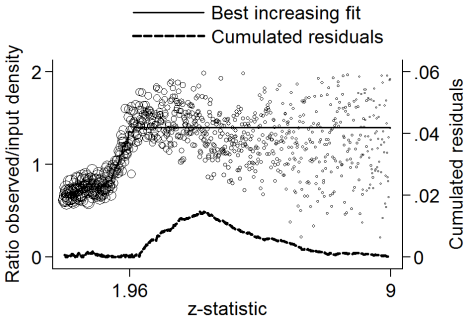
(b) Cauchy(0.5) input.



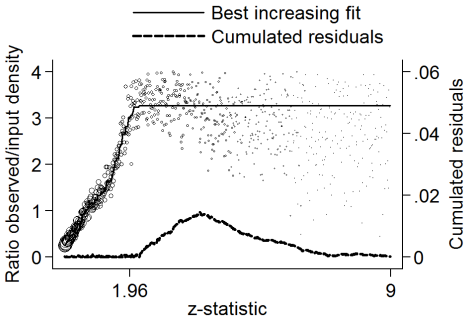
(c) WDI input.



(d) VHLSS input.



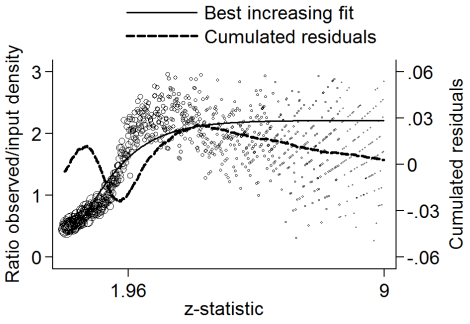
(e) QOG input.



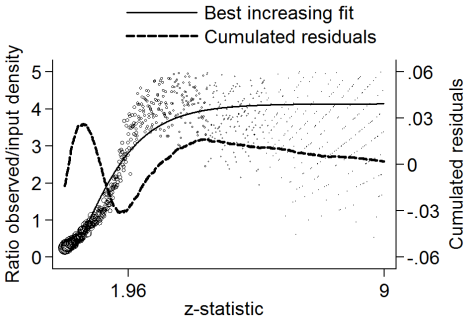
(f) PSID input.

Sources: AER, JPE, and QJE (2005-2011).

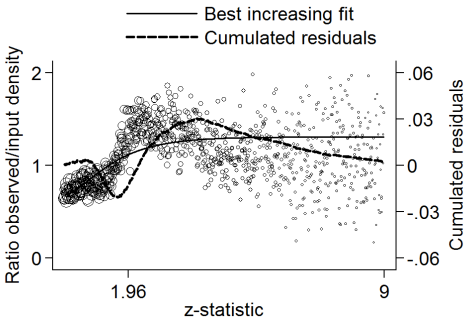
Figure 5: Parametric estimation of selection and inflation.



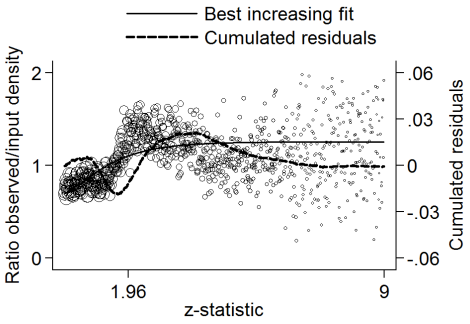
(a) Student(1) input.



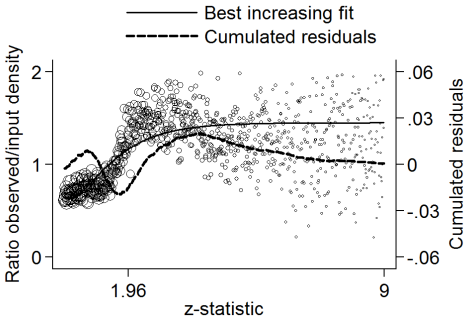
(b) Cauchy(0.5) input.



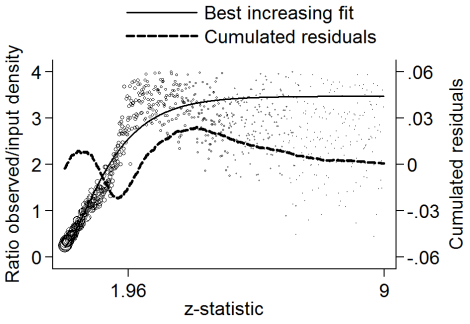
(c) WDI input.



(d) VHLSS input.



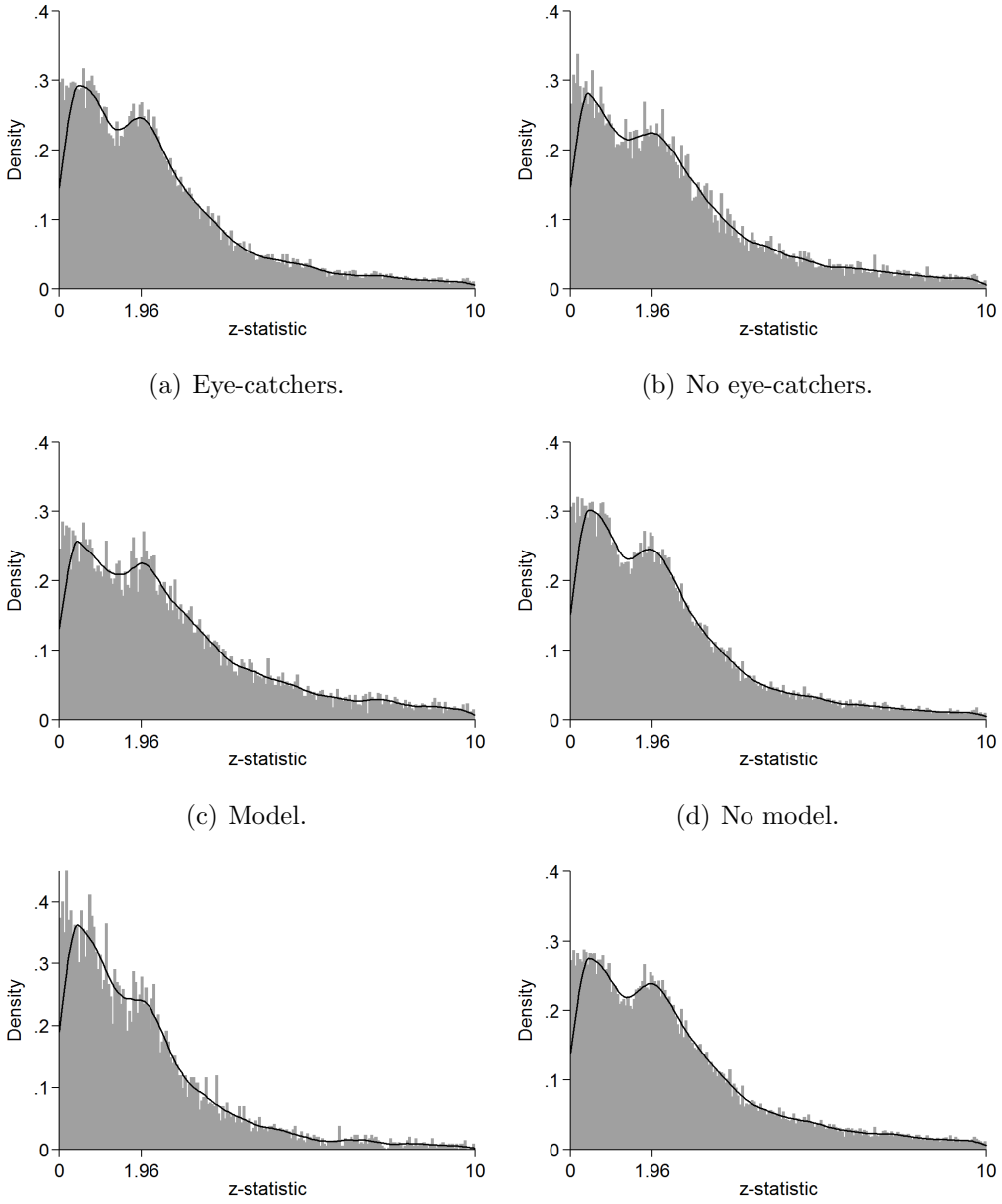
(e) QOG input.



(f) PSID input.

Sources: AER, JPE, and QJE (2005-2011).

Figure 6: Distributions of z-statistics for different sub-samples: use of eye-catchers, presence of a theoretical contribution, and type of data used.



(a) Eye-catchers.

(b) No eye-catchers.

(c) Model.

(d) No model.

(e) Lab. experiments or RCT data.

(f) Other data.

Sources: AER, JPE, and QJE (2005-2011). Distributions are unweighted and plotted using de-rounded statistics. Lines correspond to kernel density estimates.

Table 1: Descriptive statistics.

	Number of ...		
	Articles	Tables	Tests
American Economic Review	327	1,561	21,934
	[51]	[46]	[44]
Quarterly Journal of Economics	110	625	9,311
	[17]	[18]	[19]
Journal of Political Economy	204	1,203	18,833
	[32]	[35]	[38]
Macroeconomics	154	887	13,563
	[24]	[26]	[27]
Microeconomics	487	2,502	36,515
	[76]	[74]	[73]
Positive results	544	2,778	40,582
	[85]	[82]	[81]
Mixed results	81	508	8,408
	[13]	[15]	[17]
Null results	16	103	1,088
	[2]	[3]	[2]
Using eye-catchers	385	2,043	32,269
	[60]	[60]	[64]
Main results		2,487	35,288
		[73]	[70]
With model	229	979	15,727
	[36]	[29]	[31]
Single-authored	135	695	10,586
	[21]	[21]	[21]
At least one editor	400	2,145	31,649
	[62]	[63]	[63]
At least one tenured author	312	1,659	25,159
	[49]	[49]	[50]
With research assistants	361	2,009	30,578
	[56]	[59]	[61]
Data and codes available	292	1,461	20,392
	[46]	[43]	[41]
Lab. experiments or RCT data	122	593	7,535
	[19]	[17]	[15]
Other data	522	2,798	42,543
	[81]	[83]	[85]

Sources: AER, JPE, and QJE (2005-2011). This table reports the number of tests, tables, and articles for each category. *Tables* are tables or results' groups presented in the text. Proportions relatively to the total population are indicated between brackets. *Macroeconomics* and *microeconomics* are two aggregated research fields. *Positive*, *mixed*, and *negative results* correspond to the nature of the main contribution as stated by authors. *Using eyes-catchers* corresponds to articles or tables using stars or bold printing to highlight statistical significance. *Main* corresponds to results non explicitly presented as robustness checks, additional or complementary by the authors. *At least one editor* corresponds to articles with at least one member of an editorial board prior to the publication year among the authors. *At least one tenured author* corresponds to articles with at least one full professor three years before the publication year among the authors. *Data and codes available* corresponds to articles for which data and codes can be directly downloaded from the journal's website. *Lab. experiments or RCT data* stands for tests using data from laboratory experiments or randomized control trials. The sum of articles or tables by type of data slightly exceeds the total number of articles or tables as results using different data sets may be presented in the same article or table.

Table 2: Summary of parametric and non-parametric estimations using various inputs.

Maximum cumulated residual from non-parametric estimation

	De-rounded statistics			Raw statistics	
Input	Unweighted	Weighted by article	Weighted by table	Full sample	Excluding low-precision
Student(1)	0.023	0.022	0.023	0.018	0.021
Cauchy(0.5)	0.015	0.015	0.016	0.011	0.014
WDI	0.034	0.032	0.033	0.028	0.031
VHLSS	0.027	0.025	0.025	0.021	0.023
QOG	0.021	0.019	0.019	0.015	0.018
PSID	0.021	0.019	0.019	0.015	0.017

Maximum cumulated residual from parametric estimation

	De-rounded statistics			Raw statistics	
Input	Unweighted	Weighted by article	Weighted by table	Full sample	Excluding low-precision
Student(1)	0.030	0.029	0.030	0.026	0.027
Cauchy(0.5)	0.023	0.017	0.017	0.016	0.018
WDI	0.033	0.037	0.038	0.031	0.032
VHLSS	0.024	0.028	0.028	0.021	0.022
QOG	0.023	0.025	0.026	0.020	0.021
PSID	0.031	0.027	0.026	0.024	0.025

Sources: AER, JPE, QJE (2005-2011), and authors' calculation. See the text for the definitions of weights and sample restrictions.

Table 3: Summary of non-parametric and parametric estimations using the empirical WDI input for various sub-samples.

Sample	Maximum cumulated residual from non-parametric estimation	Maximum cumulated residual from parametric estimation
Macroeconomics	0.027	0.033
Microeconomics	0.030	0.030
Positive results	0.028	0.032
Null results[1]	0.073	0.007
Eye-catchers	0.036	0.037
No eye-catchers	0.015	0.020
Main results	0.024	0.028
Non-main results	0.042	0.040
With model	0.008	0.012
Without model	0.040	0.039
Low average PhD-age	0.048	0.047
High average PhD-age	0.011	0.015
No editor	0.032	0.033
At least one editor	0.026	0.030
No tenured author	0.040	0.041
At least one tenured author	0.017	0.021
Single authored	0.041	0.040
Co-authored	0.024	0.028
With research assistants	0.034	0.035
Without research assistants	0.018	0.023
Low number of thanks	0.019	0.023
High number of thanks	0.035	0.036
Data and codes available	0.029	0.032
Data or codes not available	0.027	0.030
Lab. experiments or RCT data[1]	0.053	0.041
Other data	0.024	0.029

Sources: AER, JPE, QJE (2005-2011) and authors' calculation. *Low average PhD-age* corresponds to articles written by authors whose average age since PhD is below the median of the articles' population. *Low number of thanks* corresponds to articles where the number of individuals thanked in the title's footnote is below the median of the articles' population. See notes of table 1 for the definitions of other categories.

[1]: These estimates are not reliable. In the case of articles reporting null results as their main contribution, the number of observations is way too low to apply our accounting method. In the case of laboratory experiments or randomized control trials, large z-statistics are less likely to appear which violates our methodological hypothesis that selection is increasing.

Appendix

Proof. Lemma 1.

As f is strictly increasing in e for any given z, there exists a unique h_z such that:

$$f(z,e) \geq F \Leftrightarrow e \geq h_z$$

Note that the function $h : z \mapsto h_z$ should be non-increasing. Otherwise, there would exist $z_1 < z_2$ such that $h_{z_1} < h_{z_2}$. This is absurd as $F = f(z_1, h_{z_1}) \leq f(z_2, h_{z_1}) < f(z_2, h_{z_2}) = F$. This part shows that an increasing function $\tilde{G}$ verifying $\tilde{G}(h(z)) = 1 - r(z)$ can easily be constructed and is uniquely defined on the image of h. Note that G is not uniquely defined outside of this set. This illustrates that G can take any values in the range of contributions where articles are always rejected or accepted irrespectively of their t-statistics.

Finally, we need to show that such a function $\tilde{G}$ can be defined as a surjection $(-\infty, \infty) \mapsto [0,1]$, i.e. $\tilde{G}$ can be the cumulative of a distribution. To verify this, note that on the image of h, $\tilde{G}$ is equal to $1 - r(z)$. Consequently, $\tilde{G}(h([0, T_{lim}])) \subset [0,1]$ and $\tilde{G}$ can always be completed outside of this set to be a surjection.

Note that for any given observed output and any selection function, an infinite sequence $\{G_z\}_z$ may transform the input into the output through f. The intuition is the following: for any given z, the only crucial quantity is how many ε would help pass the threshold. The shape of the distribution above or below the key quality $h(z)$ does not matter. When we limit ourselves to an invariant distribution, G is uniquely determined as $h(z)$ covers the interval of contribution. □

Proof. Corollary 1.

Given lemma 1, the only argument that needs to be made is that the image of the function $\int_0^\infty \int_0^\infty \left[1_{f(z,\varepsilon)\geq F} dG_z(\varepsilon) d\varepsilon\right] \varphi(z) dz \times \tilde{r}$ is in $[0,1]$. To prove this, remark first that the image of $\int_0^\infty \int_0^\infty \left[1_{f(z,\varepsilon)\geq F} dG_z(\varepsilon) d\varepsilon\right] \varphi(z) dz \times \psi/\varphi$ is in $[0,1]$ as it is equal to $\int_0^\infty \left[1_{f(z,\varepsilon)\geq F} dG_z(\varepsilon) d\varepsilon\right]$. Finally, note that $\max_{[0,\infty)}(f) \leq \max_{[0,\infty)}(\psi/\varphi)$ and $\min_{[0,\infty)}(f) \geq \min_{[0,\infty)}(\psi/\varphi)$. Otherwise, the function equal to $\tilde{r}$ but bounded by the bounds of ψ/φ would be a better increasing fit of the ratio ψ/φ. □