Star wars: The empirics strike back

Abel Brodeur, Mathias Lé, Marc Sangnier, Yanos Zylberberg

HAL Id: halshs-00710122

https://halshs.archives-ouvertes.fr/halshs-00710122

Submitted on 20 Jun 2012

Star wars: The empirics strike back

Abel Brodeur
Mathias Lé
Marc Sangnier
Yanos Zylberberg

STAR WARS: THE EMPIRICS STRIKE BACK[*]

Abel Brodeur[†]   Mathias Lé[‡]   Marc Sangnier[§]

Yanos Zylberberg[¶]

June 2012

Abstract

Journals favor rejections of the null hypothesis. This selection upon results 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 camel shape with abundant p-values above .25, a valley between .25 and .10 and a bump slightly under .05. Missing tests are those which would have been accepted but close to being rejected (p-values between .25 and .10). We show that this pattern corresponds to a shift in the distribution of p-values: between 10% and 20% of marginally rejected tests are misallocated. Our interpretation is that researchers might be tempted to *inflate* the value of their tests by choosing the specification that provides the highest statistics. Note that *Inflation* is larger in articles where stars are used in order to highlight statistical significance and lower in articles with theoretical models.

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

JEL CODES: A11, B41, C13, C44.

[†]Paris School of Economics; abel.brodeur@parisschoolofeconomics.eu

[‡]Paris School of Economics; mle@pse.ens.fr

[§]Sciences Po and Paris School of Economics; sangnier@pse.ens.fr

[¶]CREI (Universitat Pompeu Fabra); yzylberberg@crei.cat (corresponding author); (+34) 93 542 1145; Ramon Trias Fargas, 25-27, 08005-Barcelona, Spain

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, Melody Gardot)

The introduction of norms – confidence at 95% or 90% – coupled with the use of eye catchers – stars – has led the academic community to accept more easily starry stories with marginally significant coefficients than starless ones with marginally insignificant coefficients. As highlighted in the seminal paper of 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. First, the choice of a norm was precisely made to strongly discriminate between rejected and accepted hypotheses.[1] Second, the incapacity to reject a hypothesis might be due to weaknesses in the methodology.

As an unintended consequence, researchers may now anticipate this selection and consider that it is a stumbling block for their ideas to be considered. As such, among a set of acceptable specifications for a test, they may be tempted to keep those with the highest statistics in order to increase their chances of being published. Keeping only such specifications would lead to an *inflation* in the statistics of observed tests.[2]

---

[1] R. A. Fisher is the one who institutionalized the significance levels (Statistical Methods for Research Workers 1st Edition 1925). 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 way we understand the inflation bias is as follows. Imagine that there is only one hypothesis tested per paper. For each hypothesis is attached a set of acceptable specifications and subsequently a distribution of z-statistics. Imagine for the simplicity of exposure that journals care about the average z-statistics. In our interpretation, rejection by journals or self-censorship by authors (selection in general) consist in the *full* censorship of the distribution of z-statistics associated to this hypothesis: this distribution of z-statistics is *fully* excluded if it does not satisfy the criterion. On the other hand, inflation is a *partial* censorship of results: only a subset of the distribution of z-statistics (i.e. a subset of acceptable specifications) is shown. Consequently, the submitted and the published distributions might have different means than the set of acceptable specifications. In addition, researchers sometimes strongly believe that an effect *should* be captured by the data and specifications delivering insignificant results will be disregarded ex-post. Unconsciously, the choice of the right specification depends on its capacity to detect an effect.

This effect should be less present when the test is largely accepted/rejected.

This inflation bias should have different implications on the observed distribution of tests than selection or self-censorship. Our methodology to identify the inflation bias is the following. We argue that selection – by authors themselves, referees or journals – should be weakly increasing with the value of a statistic exhibited in a paper. Imagine now that researchers finding a set of specifications with p-values centered around .15 were to exhibit the only specification with a p-value slightly under .05. This behavior would generate shift in the observed distribution of statistics which would be inconsistent with the previous assumption on selection. There would be (i) not enough tests around .15 (as if they were disliked relatively to .30 tests) and (ii) too many slightly under .05 (as if they were better than .001 tests).

We find support for this inflation bias. The distribution of test statistics published in three of the most prestigious economic journals over the period 2005-2011 exhibits a sizeable under-representation of marginally insignificant statistics. In a nutshell, once tests are all normalized as z-statistics, the distribution has a camel shape with (i) missing z-statistics between 1.2 and 1.65 (p-values between .25 and .10) and a local minimum around 1.5 (p-value of .12), (ii) a bump between 2 and 4 (p-values below .05). We argue that this pattern cannot be explained by selection and derive a lower bound for the inflation bias under the assumption that selection should be weakly increasing in the exhibited z-statistic. We find that between 10% and 20% of tests with p-values between .05 and 0.0001 are misallocated. Note that the bulk of z-statistics echoing the missing z-statistics is at the very start of the "significant zone". Interestingly, the interval between the valley and the bulk of p-values corresponds precisely to the highest marginal returns for the selection function.[3]

---

[3]It is theoretically difficult to isolate inflation from selection: one may interpret selection and inflation as the equilibrium outcome of a game played by editors/referees and authors (Henry 2009). Editors and referees prefer to publish results that are "significant". Authors are tempted to inflate (with a cost), which pushed the editors toward being even more conservative. We believe that the observed selection and inflation are indeed exacerbated by the rational expectations of the different agents. A strong argument in favor of this race between editors/referees and authors would be an

Results vary along the subsamples of tests considered. For instance, inflation is less present 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 act as a commitment to keep inflation low. 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. Articles with theoretical models and experiments (laboratory experiments or randomized control trials) are also less prone to the inflation bias.

The literature on tests in economics was flourishing in the eighties and already pointed out some concerns evoked in this article. On the inflation bias, 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 thus suggest to display the range of inferences generated by a set of models. Leamer (1983) rule out the myth inherited from the physical sciences that econometric inferences are independent of priors. It is possible to exhibit both a positive or a negative effect of capital punishment on murder depending on priors on the acceptable specification. Lovell (1983) and Denton (1985) are close to the present paper and study the implications of individual and collective data mining. De Long and Lang (1992) already suspect that the classical distribution of z-statistics is not obtained. However, they essentially insist on a specific form of publication bias, known as the file drawer problem: z-statistics with values near zero are not likely to be published. We will refer to this file drawer issue as being part of *selection* among other mechanisms such as self-censoring of insignificant results by authors.

A large number of recent publications identify this problem (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

increasing selection even below .05, i.e. editors challenge the credibility of z-stats slightly above 2. Our findings do not show this pattern.

and positive returns to education. They implement a generalized method to identify reporting bias. This method has been developed by Hedges (1992) and extended by Doucouliagos and Stanley (2011). Card and Krueger (1995) and Doucouliagos et al. (2011) are two other examples of meta-analysis dealing with publication bias. The selection issue has also received a great deal of attention by the medical literature (Berlin et al. 1989).

To our knowledge, this project is the first to identify a residual that cannot be explained by selection and to propose a way to measure it. The major hindrance is the need for a census of tests in the literature. Identifying tests necessitates (i) a good understanding of the argument developed in an article and (ii) a strict process avoiding any subjective selection of tests. The first observation restricts the set of potential research assistants to economists and the only economists with a sufficiently low opportunity cost were ourselves. We tackle the second issue by being as conservative as possible, and by avoiding interpretations of the intentions of the author. We report all the tests discussed in the body of the article by authors. In the end, this collecting process generates 49,765 tests grouped in 3,437 tables or results subsections and 637 articles, extracted from the AER, JPE and QJE over the period 2005-2011.

Section I. details the methodology to construct the dataset, provides some descriptive statistics, and documents the raw distribution of tests. Section II. provides a theoretical framework and the associated empirical strategy. Finally, in section III., we discuss the main results and condition the analysis to different types of articles.

I. Data

In this section, we describe the reporting process of tests collected 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 and derive the raw distribution of tests. Finally, we propose some methods to alleviate the over-representation of round values and the potential overweight attributed to articles with many tests.

A.   Reporting process

The ideal measure of interest of this article is the reported value of formal tests of central hypotheses. In practice, the huge majority of those formal tests are two-sided tests for regressions' coefficients and are implicitly discussed in the body of the article (i.e. "coefficients are significant"). To simplify the exposition, we will explain the process as if we only had two-sided tests for regressions' coefficients but the description applies to our treatment of other tests. Not all coefficients reported in tables should be considered as tests of central hypotheses. Accordingly, we trust the authors and report tests that are discussed in the body of the article except if they are explicitly described as controls. The choice of this process helps to get rid of cognitive bias at the expense of parsimony. With this mechanical way of reporting tests, we also report statistical tests that the authors may expect to fail, but we do not report explicit placebo tests. Sometimes, however, the status of a test was unclear when reading the paper. In line with the methodology discussed above, we prefer to add a non-relevant test than censor a relevant one. The final dataset might include tests of controls as long as their results are extensively discussed by the author without explicit reference to them as being controls.

As we are only interested in tests of central hypotheses of articles, we also exclude descriptive statistics or groups comparisons.[4] A specific rule concerns two-stages procedures. We do not report first-stages, except if the first-stage is described by the author as a major contribution of the article. We do report tests in extensions

---

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

or robustness tests. Other rules are as follows. We report numbers exactly as they are presented in articles, i.e. we never round them up or down. Importantly, we report the issue of the journal, the starting page of the article and give the position of the test in the article (page or table, panel, row, column).

We report some additional information on each test, i.e. its type (one-sided, two-sided, Spearman correlation test, Mann-Whitney, etc.) and the status of the test in the article (main, non-main). As above, 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". Finally, the number of authors, JEL codes when available, the presence of a theoretical model, the type of data (laboratory experiment, random experiment or other) and the use of eye-catchers (stars or other formatting tricks such as bold printing associated to a table) are also recorded. We do not report the sample size and the number of variables (regressors) as this information is not always provided by authors.

B. Descriptive statistics

The previous collecting process groups 3 types of measures, p-values or t-statistics when directly reported by authors and coefficients and standard errors for the vast majority of tests. In order to get an homogeneous measure, we transform p-values into the equivalent z-statistics (1.96 would be associated to a p-value of .05). As for coefficients and standard errors, we simply construct the ratio of the two. 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 will treat the ratio of the coefficient over the standard error 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 .05.

The transformation into z-statistic allows us to observe more clearly the fat tail of tests (with small p-values). Figure I(a) presents the raw distribution. Notice that a very large number of p-values end up below the .05 threshold (more than 50% of tests are rejected at this significance level). Thus, on the 49,765 tests extracted from the 3 journals, around 30,000 are rejected at .10, 26,000 at .05, 20,000 at .01. Table I gives the decomposition of these 49,765 tests along several dimensions. The number of tests per article is surprisingly high (a bit less than 80 on average) and mainly driven by some articles with very large number of tests reported. The median article has a bit more than 50 tests distributed among 5 tables. We think these figures are reasonable. Imagine a paper with two variables of interest (i.e. democracy and institutions), six different specifications per table and 5 tables. We would report 60 coefficients, a bit more than our median article.

Journals do not contribute equally. There is an over-representation of the American Economic Review (more articles are published in the AER than in the two other journals) and an under-representation of the Journal of Political Economy. In addition to the raw number of tests, we give the number of articles and tables from which these tests have been extracted. Articles from the AER represent more than half of the total (since AER articles are shorter and with fewer tables, they represent 51% of the number of articles but only 44% of the tests).

More interestingly, less 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. Experiments constitute a small part of the overall sample. To be more precise, the AER accepts a lot of experimental articles while the QJE favors 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). With the conservative way of reporting main results, more than 70% of tables/results from which tests are extracted are considered as main.

90% of tests are two-sided tests of a coefficient. Stars are used for three quarters of those tests. Most of the times, the starting threshold is .10 rather than .05. We define henceforth the use of eye-catchers as the use of stars and bold in a table, excluding the explicit display of p-values.

C. The distribution of tests

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, a small proportion of coefficients and standard deviations are reported with a pretty poor precision (0.020 and 0.010 for example). Reconstructed z-statistics are thus over-abundant for fractions of integers ($\frac{1}{1}, \frac{2}{1}, \frac{3}{1}, \frac{1}{3}, \frac{1}{2} \dots$). Second, some authors report a lot of versions of the same test. In some articles, more than 100 values are reported against 4 or 5 in others. 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.[5]

To alleviate the first issue, we randomly redraw a value in the interval of potential z-statistics given the reported values and their precision. In the example given above, the interval would be $[\frac{0.0195}{0.0105}, \frac{0.0205}{0.0095}] \approx [1.86, 2.16]$. We draw a z-statistic from a uniform distribution over the interval and replace the previous one. This reallocation should not have any impact on the analysis other than smoothing potential discontinuities.[6]

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

---

[5]For example, one might conjecture that authors report more tests when the significance of those is shaky. Conversely, one may also choose to display a small number of satisfying tests as others would fail.

[6]For statistics close to significance levels, we could take advantage of the information embedded in the presence of a star. However, this approach could only be used for a reduced number of observations, and for tables where stars are used.

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 subsection) 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 I(b) presents the unweighted distribution. The shape is striking. The distribution presents a camel pattern with a local minimum around $z = 1.5$ (p-value of .12) and a local maximum around z equals 2 (p-value under .05). The presence of a local maximum around 2 is not very surprising, the existence of a valley before more so. Intuitively, selection could explain an increasing pattern for the distribution of z-statistics at the beginning of the interval $[0, \infty)$. On the other hand, it is likely that there is a natural decreasing pattern of the distribution over the whole interval. Both effects put together could explain the presence of a unique local maximum, a local minimum less so. Our empirical strategy will consist in formalizing this argument: only a shift of statistics can generate such a pattern and the inflation bias seems to explain this shift.[7]

Figures I(c) and (d) present the weighted distributions. The camel shape is more pronounced than for the unweighted distribution. 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 conservative way to report tests might have included tests of not-central hypotheses.

Figure II presents the unweighted distributions of z-statistics over various sub-samples of articles. II(a) and (b) are decomposition along the use of eye-catchers,

---

[7]In our web Appendix, we also test for discontinuities. We find evidence that the total distribution of tests presents a small discontinuity around the threshold .10, not much around the .05 or the .01 thresholds. This modest effect might be explained by the dilution of hypothesis testing in journal articles. In the absence of a single test, empirical economists provide many converging arguments under the form of different specifications or different samples for a single effect. Besides, an empirical article is often dedicated to the identification of more than one mechanism. As such, the real z-statistic related to an article is a distribution or a set of arguments and this dilution smoothes potential discrepancies around thresholds.

II(d) and II(d) along the presence of a theoretical model. The web appendix also provides the distributions for each journal and each year and discriminate between laboratory experiments/randomized control trials and other types of experimental settings. The camel shape is more pronounced in tables where authors choose to use eye catchers, in articles that do not include any explicit theoretical contribution and for non-RCT/experimental papers. For the last category, test statistics are globally lower, potentially highlighting lower selection.

Finally, one may think that there exists two natural modes in the distribution of z-statistics. For example, for macroeconomic studies, with fewer observations and low precision, the mode would be 0. Hypotheses may be generally accepted because of the lack of precision. For microeconomic studies, with a better precision, hypotheses may be rejected more easily and the mode would be around 2. Aggregating these two types of studies would generate the pattern that we observe. However, the shape that we uncover is very robust to decompositions along JEL code categories.

II. Theoretical framework and empirical strategies

In this section, we present our estimation strategy. The idea is that the observed distribution of z-statistics may be thought as generated by (i) an input, (ii) a selection function over results, and (iii) a noise, which will partly capture inflation.

We first present a very simple model of selection in academic publishing. In this framework, under the assumption that selection favors high over low statistics, the ratio of the observed density of z-statistics over the input is increasing in z. The empirical strategy will consist in capturing any violation of this prediction and relate it with the inflation bias. Finally, we discuss the range of distributions chosen as inputs and stories that may challenge our interpretation.

A. The selection process

We consider a very simple theoretical framework of selection into journals. We abstract from authors and directly consider the universe of working paper.[8] Each economic working paper has a unique hypothesis which is tested with a unique specification. Denote z the absolute value of the statistic associated to this test and φ 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 is a noise[9] entering into the selection process. 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 high ε correspond to articles with higher likelihood to be published, for a same z. Denote G_z the distribution of ε conditional on the value of z.

The density of tests 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 for a given z depends on the quantity of articles with ε sufficiently high to pass the threshold and on the input. In the black box which generates the output from the input, two effects intervene. First, as the value of z changes, the minimum noise ε required to pass the threshold changes. This is the selection effect. Second, the distribution G_z of this ε might change conditionally on z.

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

This assumption implies that if we shut down the second channel, i.e. the noise

---

[8]Note that the selection of potential economic issues into a working paper is not modeled here. You 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.

[9]We label here ε as a noise but it can also capture inclinations of journals for certain articles, the importance of the question, the originality of the methodology, or the quality of the paper.

is independent of z, then the ratio $\psi(z)/\varphi(z)$ is (weakly) increasing in z.

This assumption that journals, referees and authors prefer tests rejecting the null may not be viable for high z-statistics. Such results could indicate an empirical mis-specification to referees, but 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.[10] We discuss the other potential mechanisms challenging this assumption at the end of this section.

B.   Identification strategy

The identification strategy relies on the result that, with an independent noise ε, we should see an increasing pattern in the selection process, i.e. the proportion of articles selected among the submitted/written ones should be (weakly) increasing in z. We can not explain stagnation or slowdowns in this ratio (accepted articles over the input) with selection or self-censoring alone, i.e. with a distribution of noise invariant in z. Our empirical strategy consists in estimating how well selection might explain the observed pattern and interpret the residual as violation of the independence of the noise. This strategy is conservative as it attributes all purely increasing patterns in z to selection.

Let us assume that we know the distribution φ. The ratio of the output density to the input density can be written as:

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

In this framework, once cleared from the input, the output is a function of the selection function f and the conditional distribution of noise G_z. We will isolate

---

[10]Journals and authors may privilege p-values very close to 1 and very close to 0, which would fit the camel pattern with two bumps. There is no way to formally reject this interpretation. However, we think that this effect is marginal as the number of articles for which the central hypothesis is accepted is very small in our sample.

selection $z \mapsto \mathbb{1}_{f(z,\varepsilon)}$ from noise $G_z(\varepsilon)$. We argue that this noise captures – among other potential mechanisms – local shifts in the distribution. An inflation bias corresponds to such a shift. In this framework, this would translate into productions of low ε just below the threshold against very high ε above.

As explained above, under the assumption that the selection function is increasing and with a distribution of noise invariant in z, the ratio output/input should be increasing. In fact, the reciprocal is also true: any increasing pattern for the ratio output/input can be explained by selection alone (i.e. with a distribution of noise invariant in z). Given any selection process f verifying assumption 1, any increasing function of z (in a reasonable interval) for the ratio of densities can be generated by f and a certain distribution of noise, invariant in z. Intuitively, there is no way to identify an inflation effect with an increasing ratio of densities, as an invariant distribution of noise can always be considered to fit the pattern.

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

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

$\tilde{G}$ 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. □

Following this lemma, the empirical strategy will consist in the estimation of the best-fitting increasing function $\tilde{f}$ for the ratio $\psi(z)/\varphi(z)$.

We will find the weakly increasing $\tilde{f}$ that minimize the weighted distance with the ratio ψ/φ:

$$\sum_i \left[\psi(z_i)/\varphi(z_i) - \tilde{f}(z_i) \right]^2 \varphi(z_i)$$

In order to estimate our effects, we have focused on the ratio ψ/φ. The following corollary transforms the estimated ratio in a cumulative distribution of z-statistics and relates the residual of the previous estimation to the distance between an invariant distribution of contribution (selection alone) and the observed distribution of contribution.

Corollary 1 (Residual). *Following the previous lemma, there exists a cumulative distribution $\tilde{G}$ which represents $\tilde{f}$ uniquely defined on $\{\varepsilon, \exists z \in [0, T_{lim}], f(z, \varepsilon) = F\}$, such that:*

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

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

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

where h is defined as $f(z, \varepsilon) \geq F \Leftrightarrow \varepsilon \geq h(z)$. Define $\tilde{\psi}(z) = (1 - \tilde{G}(h(z)))\varphi(z)$ the density of z-statistics associated to $\tilde{G}$, then the cumulated residual is simply

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

Proof. In the appendix. $\square$

This corollary allows us to map the **cumulated residual** of the estimation with a quantity that can be interpreted. Indeed, given z, $\int_0^z \psi(\tau)d\tau - \int_0^z \tilde{\psi}(\tau)d\tau$ is the number of z-statistics between $[0, z]$ that can not be explained by an increasing selection function.

C. Input

In practice, a difficulty arises. What do we want to consider as the exogenous input and what do we want to include in the selection process? In the process that intervenes before publication, there are several choices that may change the distribution of tests: the choice of the research question, the dataset, the decision to submit and the acceptance of referees. We think that all these processes are very likely to verify the assumption 1 (at least for z-statistics not extremely high) and that the input can be taken as the distribution before all these choices. All the choices (research question, dataset, decision to create a working paper, submission and acceptance) will thus be included in the endogenous process.

The process through which nature generates mechanisms is unobserved, the shape of the input is unknown. The idea here is to consider a large range of distributions. The classes of distribution should ideally be ratio distributions as 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 the large number of rejected tests). Let us 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 done 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 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. From the descriptive statistics, we know that selection should be quite drastic when we consider a normal distribution for the exogenous input. The output

displays more than 50% of accepted 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.

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, Cauchy distribution allows us to extend the input to fat-tail distributions.

Our last approach consists in creating an empirical counterfactual distribution of statistics obtained by random tests performed on large classic datasets.

Class 3 (Empirical). *We randomly draw 4 variables from the World Development Indicators (WDI) and run* $2,000,000$ *regressions between these variables and stock the z-statistic behind the first variable.*[11] *Other datasets/sample can be considered and the shapes are very close.*

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

Figure III(a) shows how poorly the normal distribution fits the observed one. The assumption that input comes from uncorrelated processes can only be reconciled with the observed output with a drastic selection (which would create the observed fat tail from a gaussian tail). The fit is slightly better for the standard Cauchy distribution, e.g. the Student distribution of degree 1. The proportion of accepted tests is much higher then with 44% of accepted tests at .05 and 35% at .01. Cauchy distributions

---

[11]To be consistent with the literature, we just ran two million regressions (Sala-i Martin (1997);Hendry and Krolzig (2004)).

centered in 0 and the empirical counterfactuals of statistics obtained by random tests performed on classic datasets such as the World Development Indicators have fairly similar shapes. Figure III(c) show that the Cauchy distributions as well as the WDI placebo may help to capture the fat tail of the observed distribution. III(c) focuses on the tail: Cauchy distributions with parameters between 0.5 and 2 as well as the empirical placebo fit very well the tail of the observed distribution. More than the levels of the densities, it is their evolution which gives support to the use of these distributions as input: if we suppose that there is no selection once passed a certain threshold (p<0.000001 for these levels), we should observe a constant ratio output/input.

We will consider as exogenous inputs:

- the WDI counterfactual (called hereafter placebo) which will be the higher bound in terms of fat-tail (it is close to a Cauchy of parameter 1.5),

- the Cauchy(1)/Student(1) distribution,

- and a rather thin-tail distribution, i.e. the Cauchy distribution of parameter 0.5.

These distributions cover a large spectrum of shapes.[12] We will show that results are not sensitive to changes in the choice of inputs (in the range of Cauchy distributions with parameters between 0.5 and 2).

D.   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 selection. Our interpretation is that it will capture the *local* shift of z-statistics. In addition, this quantity is a lower bound of inflation as any *globally*

---

[12]Cauchy distributions with parameters above 2 would exhibit tails that are too fat to explain the observed distribution of tests.

increasing pattern (in z) in the inflation mechanism would be captured as part of the selection effect.

Several remarks may challenge our interpretation. First, the noise ε actually includes the quality of a paper and quality may be decreasing in z. The amount of efforts put in a paper might end up being lower with very low p-values or p-values around .15. Authors might for instance misestimate selection by journals and produce low effort in certain ranges of z-statistics. Second, the selection function may not be increasing as a well-estimated zero might be interesting and there are no formal tests to exclude this interpretation. We do not present strong evidence against this mechanism. However, two observations make us confident that this preference for well-estimated zero does not drive the whole camel shape. The first argument is based on anecdotal evidence; very few papers of the sample present a well-estimated zero as their central result. Second, these preferences should not depend on whether stars are used or whether a theoretical model is attached to the empirical analysis and we find disparities along those characteristics.

In addition, imagine that authors could predict exactly where tests will end up and decide to invest in the working paper 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 always more likely to be undertaken. We may think of a very simple setting where it is unlikely to be the case: when designing experiments (or RCTs), researchers compute power tests such as to derive the minimum number of participants for which an effect can be statistically captured. The reason is that experiments are expensive and costs need to be minimized under the condition that a test may settle whether the hypothesis is true or not. We should expect a thinner tail for those experimental settings (and this is exactly what we observe). In the other cases, the limited capacity of the author to predict where the z-statistics may end up as well as the modest incentives to limit oneself to small samples make it more implausible.

III.   Main results

This section presents the empirical strategy tested on the whole sample of tests and on subsamples. A different estimation of the best-fitting selection function will be permitted for each subsample, as the intensity of selection may differ for theoretical papers or RCT.

A.   Non-parametric estimation

For any sample, we group observed z-statistics by bandwidth of .01 and limit our study to the interval $[0, 10]$. Accordingly, the analysis is made on 1000 bins. As the empirical input appears in the denominator of the ratio $\psi(z)/\varphi(z)$, we smooth it with an Epanechnikov kernel function and a bandwidth of 0.1 in order to dilute noise (for high z, the probability to see an empty bin is not negligible).

Figures IV(a) and IV(b) give both the best increasing fit for the ratio output/placebo input and the partial sum of residuals, i.e. our lower bound for the inflation bias.[13] Results are computed with the Pool-Adjacent-Violators Algorithm.[14]

Two interpretations emerge from this estimation. First, the best increasing fit displays high marginal returns to the value of statistics only for $z \in [1.5, 2]$, and a plateau thereafter. Selection is intense precisely where it is supposed to be discriminatory, i.e. before the thresholds. Second, the misallocation of z-statistics starts to increase slightly before $z = 2$ up to 4 (the bulk between p-values of .05 and 0.0001 can not be explained by an increasing selection process alone). At the maximum, the misallocation reaches 0.025, which means that 2.5% of the total number of t-statistics are misallocated between 0 and 4. As there is no difference between 0 and 2, we compare this 2.5% to the total proportion of z-statistics between 2 and 4, i.e.

---

[13]Note that there are less and less z-statistics per bins of width 0.01. On the right-hand part of the figure, 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.

[14]Results are invariant to the use of other algorithms of isotonic optimization.

30%. The conditional probability of being misallocated for a z-statistic between 2 and 4 is thus around 8%.

Note that the attribution of misallocated z-statistics is a bit surprising: the surplus observed between 2 and 4 is here compensated by a deficit after 4 while we should expect such a deficit before 2. This result comes from the conservative hypothesis that the pattern observed in the ratio of densities should be attributed to the selection function as long as it is increasing. Accordingly, the stagnation in this ratio which is observed before 1.7 is only accounted in the selection function. Nonetheless, as the missing tests fall in the bulk between 2 and 4, it still allows us to identify a strong shift of z-statistics. As shown by figures IV(c), (d), (e), (f), results do not change when the input distribution is approximated by a Student distribution of degree 1 (standard Cauchy distribution) and a Cauchy distribution of parameter 0.5. The results are very similar both in terms of shape and magnitude.

As already suggested by the shapes of weighted distributions (figures I(c) and (d)), the results are much stronger when the distribution of observed z-statistics is corrected such that each article contributes the same to the overall distribution (see figure V). The shape of misallocation is similar but the magnitude is approximately twice as large as in the case without weights: the conditional probability of being misallocated for a z-statistic between 2 and 4 is there between 15% and 20%. In a way, the weights may compensate for our very conservative reporting process.

A concern in this estimation strategy is that the misallocation can really reflect different levels of quality between articles with z-statistics between 2 and 4 compared to the rest. We can not exclude this possibility. Two observations however gives support to our interpretation: the start of the misallocation is right after (i) the first significance thresholds, and (ii) the zone where the marginal returns of the selection function are the highest.[15]

Finally, what do we find when we split the sample into subsamples? This analysis

---

[15]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.

might be hard to interpret as there might be selection into the different subsamples. Authors with shaky results might prefer not to use eye-catchers. Besides, papers with a theoretical model may put less emphasis on empirical results and the expected publication bias may be lower. Still, the analysis on the eye-catchers sample shows that misallocated t-statistics between 0 and 4 account for more than 3% of the total number against 1% for the no-eye catchers sample (see figure VI). The conditional probability of being misallocated for a z-statistic between 2 and 4 is around 12% in the eye-catchers sample against 4% in the no-eye-catchers one. We repeat the same exercise on the subsamples model/no model and main/not main. There seems to be no inflation in articles with a theoretical model, maybe because the main contribution of the paper is then divided between theory and empirics. The main/no main analysis is at first glance more surprising: the misallocation is slightly higher in tables that we report as not being "main" (robustness, secondary hypothesis or side results). A reason might be that robustness checks may be requested by referees when the tests are very close to the threshold. To conclude this subsample analysis, experiments exhibit a very smooth pattern: there are no real bump around .05. However, z-statistics seem to disappear after the .05 threshold. An interpretation might be that experiments are designed such as to minimize the costs while being able to detect an effect. Very large z-statistics are thus less likely to appear (which violates our hypothesis that selection is increasing).

B.   Parametric estimation

A concern of the previous analysis is that it attributes misallocated tests between 2 and 4 to missing tests after this bulk. The mere observation of the distribution of tests does not give the same impression. 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 more parametric and less conservative test by estimating the selection function

under the assumption that it should belong to a set of parametric functions.

Assume here that the selection function can be approached by an exponential polynomial function, i.e. we consider the functions $\{f, \exists\{a_i\}, f(z) = c + exp(a_0 + a_1 z + a_2 z^2)\}$. This pattern allows us to account for the concave pattern of the observed ratio of densities.[16]

Figure VII gives both the best parametric fit and the partial sum of residuals as in the non-parametric case. Contrary to the non-parametric case, the misallocation of t-statistics starts after z=1 (p-values around 30%) and is decreasing up to z=1.65 (p-values equals to 10% and first significance threshold). These missing statistics are then completely retrieved between 1.65 and 3-4. Remark that the size of misallocation is very similar to the non-parametric case.

IV. Conclusion

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

In this paper, we have identified an inflation bias in tests reported in some of the most respected academic journals in economics. 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 very conservative collecting and estimating processes. The results presented in this paper may have potentially different implications for the academic community than the already known publication bias. Even though it is unclear whether these biases should be larger or smaller in other journals and disciplines,[17] it raises questions about the importance given to values of tests *per se*.

A limit of our work is that it does not say anything about how researchers inflate results. Nor does it say anything about the importance of expectations of

---

[16]The analysis can be made with simple polynomial functions: this worsen slightly the fit.

[17]Auspurg and Hinz (2011) and Gerber et al. (2010) collect distributions of tests in journals of sociology and political science.

authors/referees/editors in the amplitude of selection and inflation. 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 (Mahoney (1977)).

Propositions have already been made in order to reduce selection and inflation biases (see Weiss and Wagner (2011) for a review). First, some 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. Second, attempts to reduce data mining have been proposed in medicine or psychological science. There is a pressure for researchers to submit their methodology/empirical specifications before running the experiment (especially because the experiment can not be reproduced). Some research grants ask researchers to submit their strategy/specifications beforehand (sample size of the treatment group for instance) before starting a study. It seems however that researchers pass through this hurdle by (1) investigating an issue, (2) applying ex-post for a grant for this project, (3) funding the next project with the funds given for the previous one.

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 Statistics* **231**(5 - 6), 636 – 660.

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), pp. 381–392.

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

De Long, J. B. and Lang, K.: 1992, Are all economic hypotheses false?, *Journal of Political Economy* **100**(6), pp. 1257–1272.

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

Doucouliagos, C. and Stanley, T. D.: 2011, Are all economic facts greatly exaggerated? theory competition and selectivity, *Journal of Economic Surveys* pp. no–no.

Doucouliagos, C., Stanley, T. and Giles, M.: 2011, Are estimates of the value of a statistical life exaggerated?, *Journal of Health Economics* (0), –.

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.

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

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), pp. 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.

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

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.

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.

Table I: Descriptive statistics.

| | | Number of ... | | |
Sample		Tests	Articles	Tables
Full		49,765	637	3,437
By journal	AER	21,226	323	1,547
		[43.28]	[50.71]	[44.54]
	JPE	9,287	110	723
		[18.93]	[17.27]	[20.82]
	QJE	18,534	204	1,203
		[37.79]	[32.03]	[34.64]
By theoretical contrib.	With model	15,502	230	977
		[31.15]	[36.11]	[28.13]
By type of data	Lab. exp.	3,503	86	343
		[7.04]	[13.50]	[9.98]
	RCT	4,032	37	249
		[8.10]	[5.81]	[7.24]
	Other	42,23	519	2,883
		[84.86]	[81.47]	[83.88]
By status of result	Main	35,108		2,472
		[70.55]		[71.18]
By use of eye catchers	Stars	32,221	383	2,141
		[64.75]	[60.12]	[61.68]

Sources: AER, JPE, and QJE (2005-2011). This table reports the number of tests, articles, and tables for each category. Proportions relatively to the total number are indicated between brackets. 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 presented in the same article or table. "Theoretical contrib." stands for "theoretical contribution". "Lab. exp." stands for "laboratory experiments". "RCT" stands for "randomized control trials".

Figure I: Distributions of z-statistics.
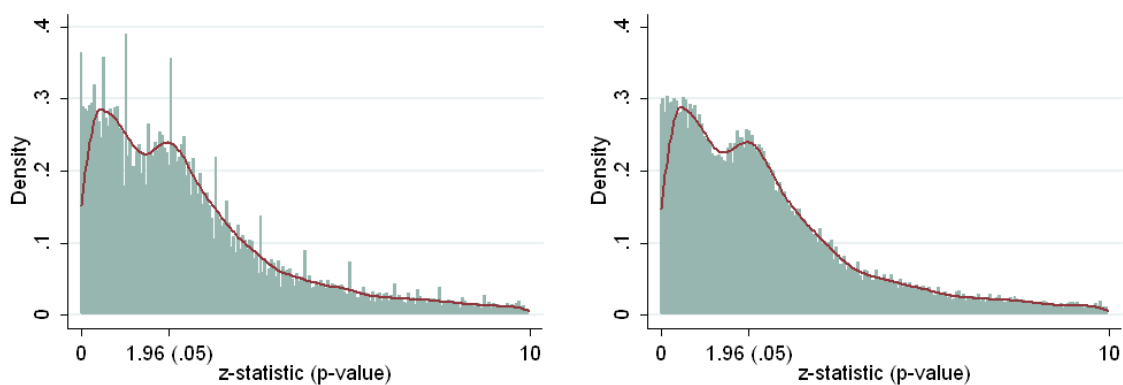
(a) Raw distribution of z-statistics.

(b) Unrounded distribution of z-statistics.
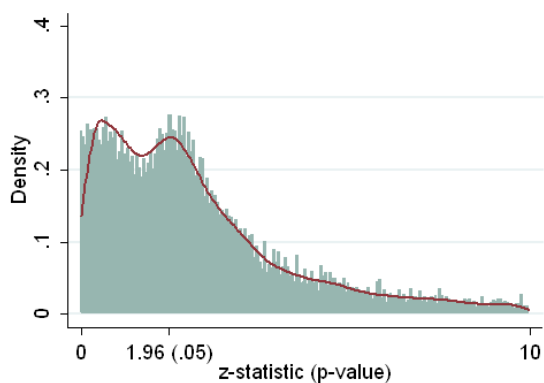
(c) Unrounded distribution of z-statistics, weighted by articles.
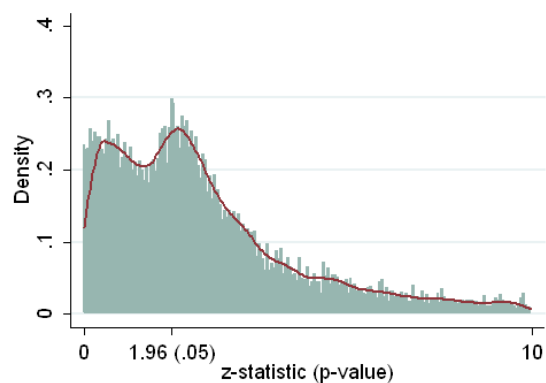
(d) Unrounded distribution of z-statistics, weighted by articles and tables.

Sources: AER, JPE, and QJE (2005-2011). See the text for unrounding 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 II: Distributions of z-statistics for different sub-samples: eyes-catchers, theoretical contribution and lab/rct experiments.
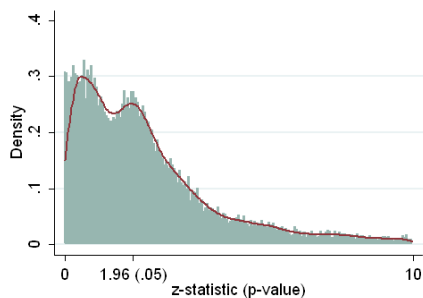
(a) Distribution of z-statistics when eyes-catchers are used.
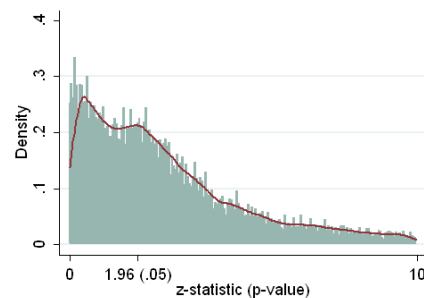
(b) Distribution of z-statistics when eyes-catchers are not used.
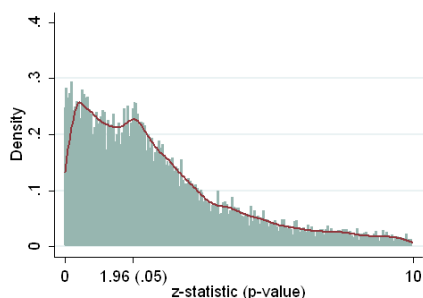
(c) Distribution of z-statistics when the article includes a model.
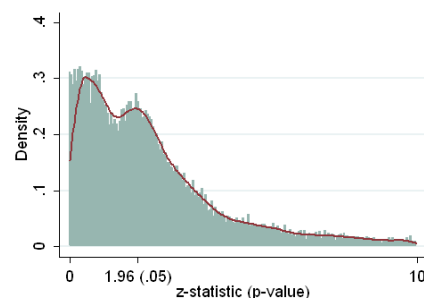
(d) Distribution of z-statistics when the article does not include a model.
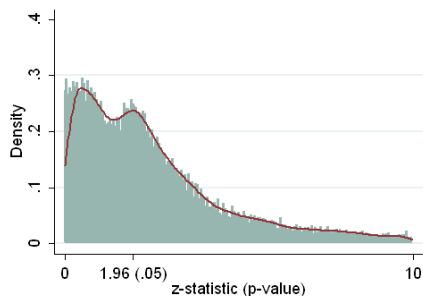
(e) Distribution of z-statistics for main tables.
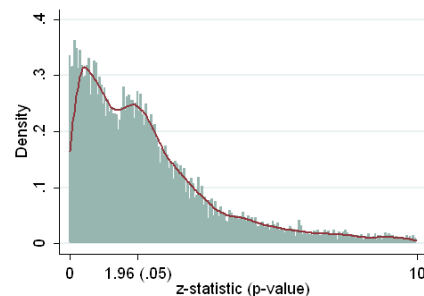
(f) Distribution of z-statistics for non-main tables.
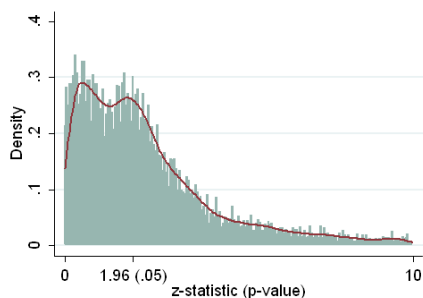
(g) Distribution of z-statistics for single-authored papers.
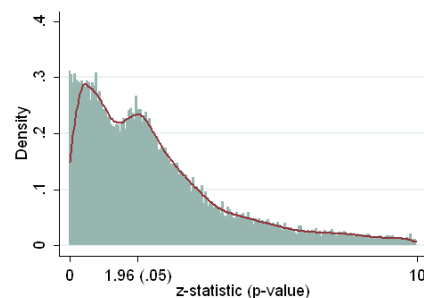
(h) Distribution of z-statistics for co-authored papers.

Figure III: Unweighted and weighted distributions of the universe of z-statistics and exogenous inputs.
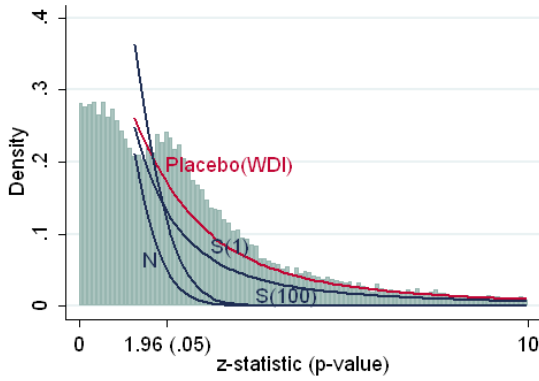
(a) Gaussian/Student inputs (0<z<10, unweighted).
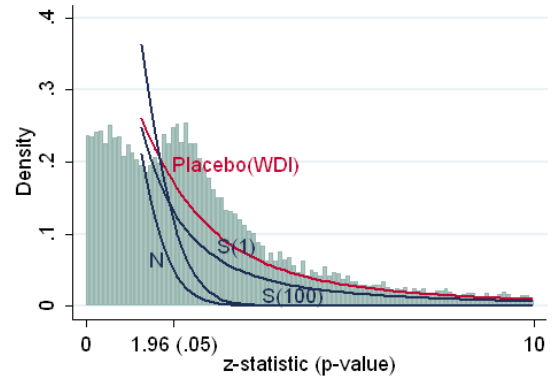
(b) Gaussian/Student inputs (0<z<10, weighted by articles).
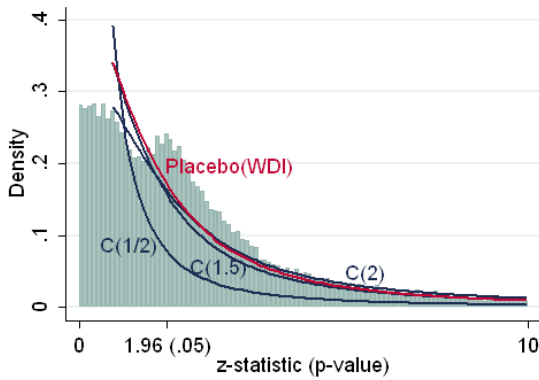
(c) Cauchy inputs (0<z<10, unweighted).
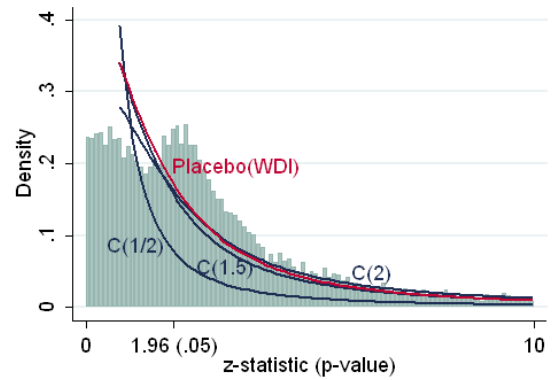
(d) Cauchy inputs (0<z<10, weighted by articles).
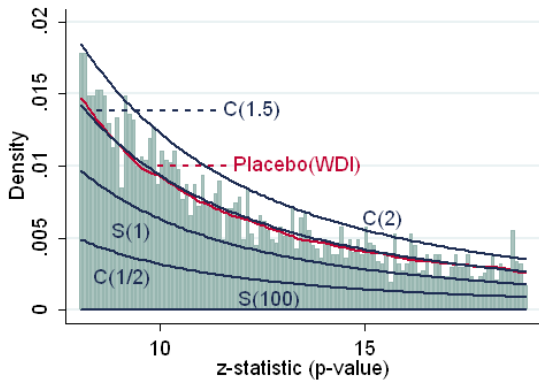
(e) All inputs (5<z<20, unweighted).
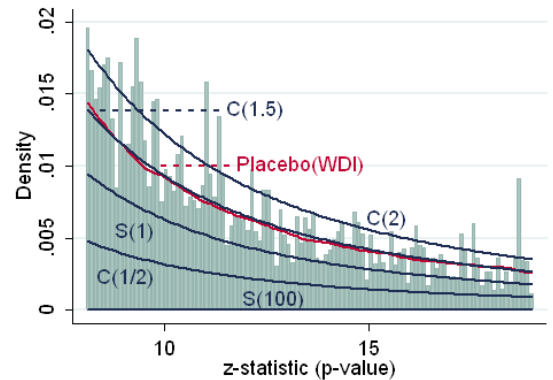
(f) All inputs (5<z<20, weighted by articles).

Sources: AER, JPE, and QJE (2005-2011). Distributions are plotted using unrounded statistics.

31

Figure IV: Non-parametric estimation of selection and inflation.
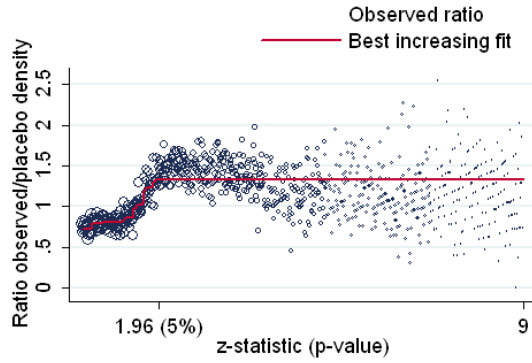
(a) Best increasing non-parametric fit for the ratio of densities output/WDI placebo input.
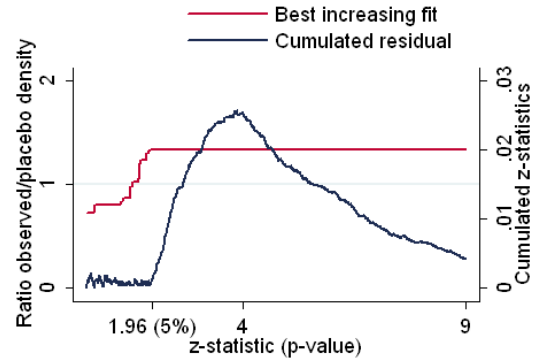
(b) Cumulated residual (WDI placebo input).
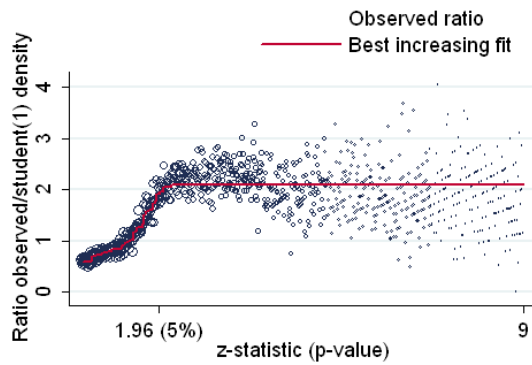
(c) Best increasing non-parametric fit for the ratio of densities output/Student input.
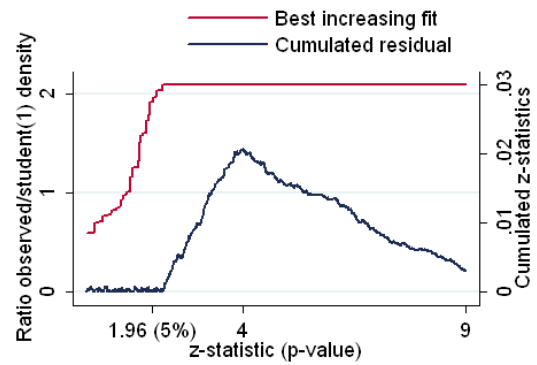
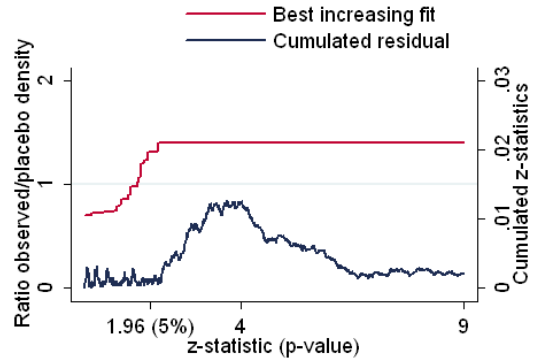(d) Cumulated residual (Student input).
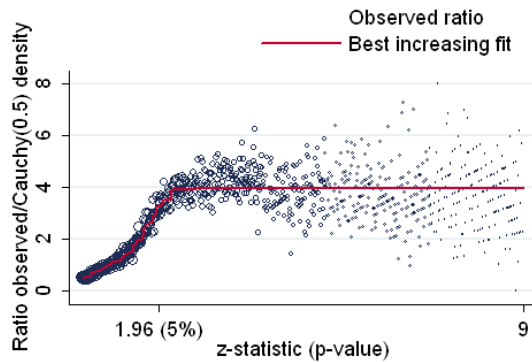
(e) Best increasing non-parametric fit for the ratio of densities output/Cauchy(0.5) input.
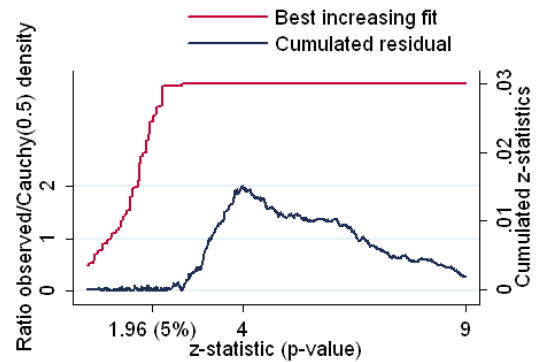
(f) Cumulated residual (Cauchy(0.5) input).

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

Figure V: Non-parametric estimation of selection and inflation (weighted distributions).
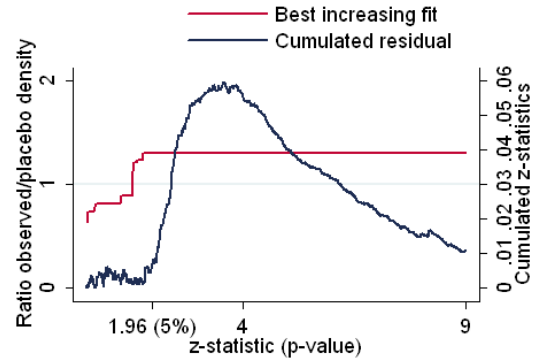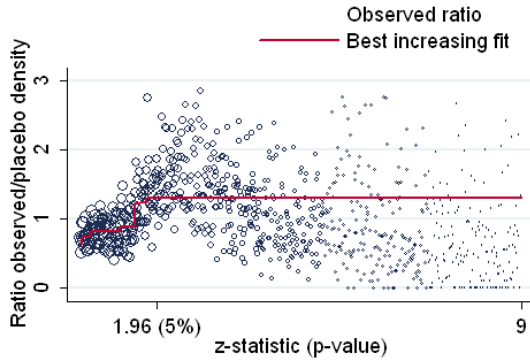
(a) Best increasing non-parametric fit for the ratio of densities output (weights articles)/WDI placebo input.

(b) Cumulated residual (WDI placebo input).
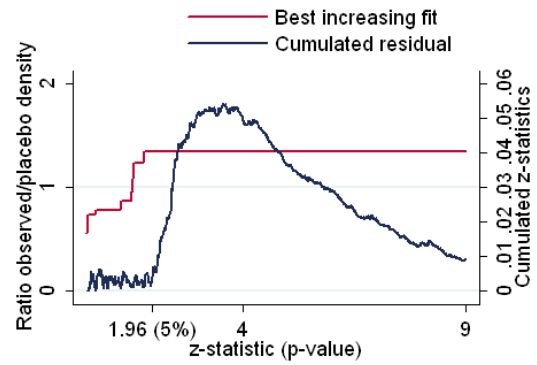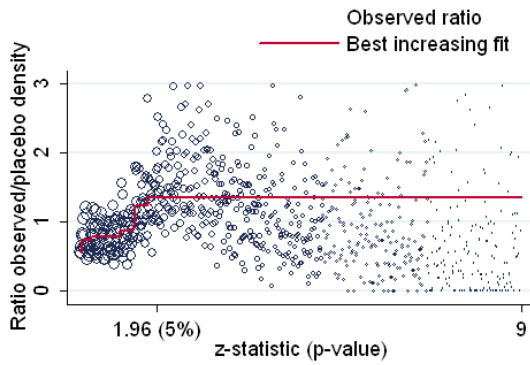
(c) Best increasing non-parametric fit for the ratio of densities output (weights tables)/WDI placebo input.

(d) Cumulated residual (WDI placebo input).

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

Figure VI: Estimation of selection and inflation for different sub-samples: eyes-catchers, theoretical contribution and lab/rct experiments.
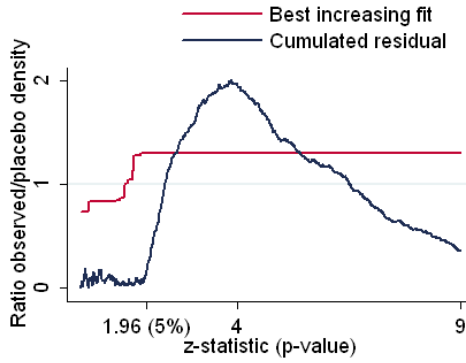
(a) Eye-catcher sample.

(b) No eye-catcher sample.
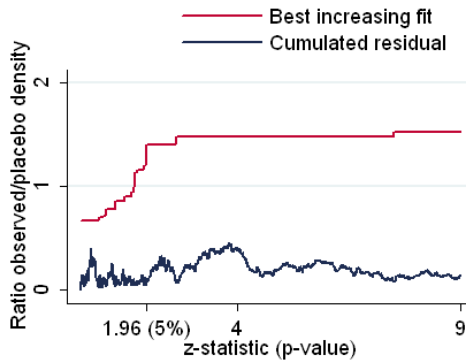
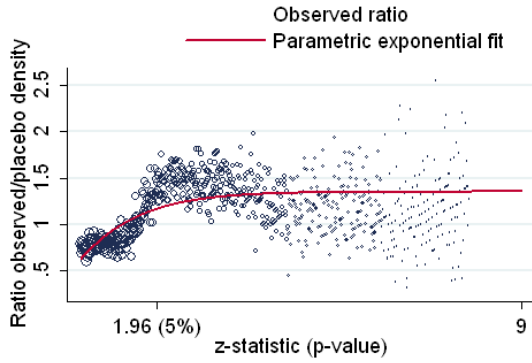(c) Theoretical framework.
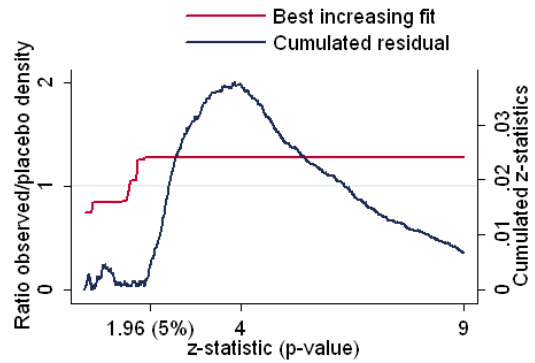
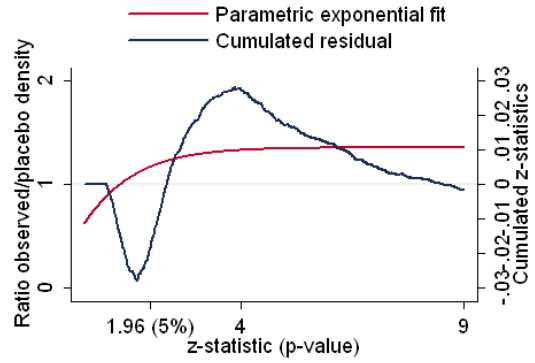(d) No theoretical framework.
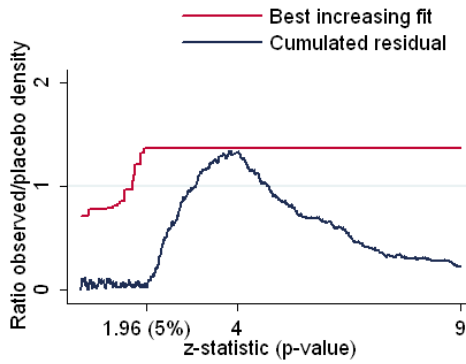
(e) Main tables.
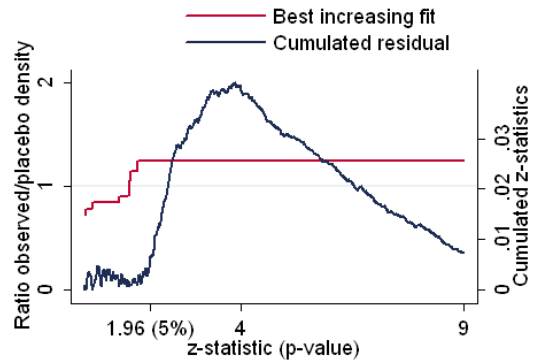
(f) No-main tables.
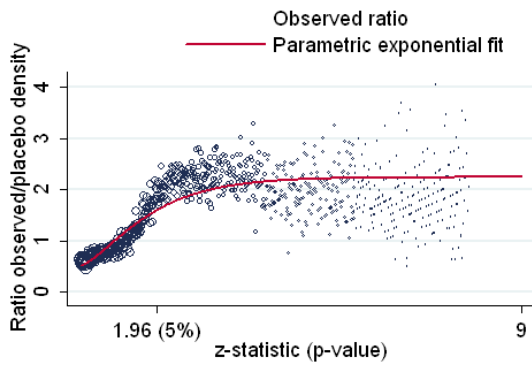
Sources: AER, JPE, and QJE (2005-2011).

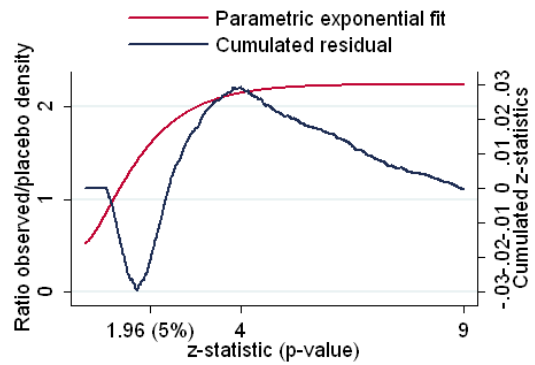Figure VII: Parametric estimation of selection and inflation.

(a) Best increasing parametric fit for the ratio of densities output/WDI placebo input.
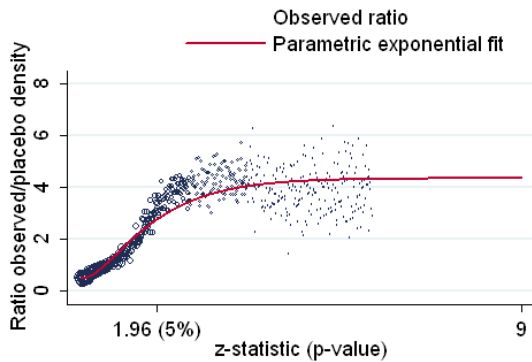
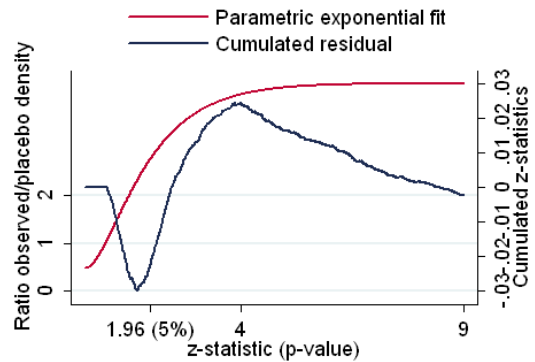(b) Cumulated residual (WDI placebo input).

(c) Best increasing parametric fit for the ratio of densities output/Student input.

(d) Cumulated residual (Student input).

(e) Best increasing parametric fit for the ratio of densities output/Cauchy(0.5) input.

(f) Cumulated residual (Cauchy(0.5) input).

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

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 $\tilde{G}(h(z)) = 1 - g(z)$ 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 $1 - g(z)$. 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. $\square$

Proof. Corollary 1.

Given lemma 1, the only argument that needs to be made is that the image of the function $\int_0^\infty \int_0^\infty \left[1_{f(z,\varepsilon) \geq F} dG_z(\varepsilon) d\varepsilon \right] \varphi(z) dz \times \tilde{f}$ 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 $\tilde{f}$ but bounded by the bounds of ψ/φ would be a better increasing fit of the ratio ψ/φ. $\qquad\square$