# Sniff Tests in Economics: Aggregate Distribution of Their Probability Values and Implications for Publication Bias

Christopher Snyder
*Dartmouth College and National Bureau of Economic Research*

Ran Zhuo
*Harvard University*

September 2018

**Abstract:** The increasing demand for rigor in empirical economics has led to the growing use of auxiliary tests (balance, specification, over-identification, placebo, etc.) supporting the credibility of a paper's main results. We dub these "sniff tests" because standards for passing are subjective and rejection is bad news for the author. Sniff tests offer a new window into publication bias since authors prefer them to be insignificant, the reverse of standard statistical tests. Collecting a sample of nearly 30,000 sniff tests across 60 economics journals, we provide the first estimate of their aggregate probability-value (p-value) distribution. For the subsample of balance tests in randomized controlled trials (for which the distribution of p-values is known to be uniform absent publication bias, allowing reduced-form methods to be employed) estimates suggest that 45% of failed tests remain in the "file drawer" rather than being published. For the remaining sample with an unknown distribution of p-values, structural estimates suggest an even larger file-drawer problem, as high as 91%. Fewer significant sniff tests show up in top-tier journals, smaller tables, and more recent articles. We find no evidence of author manipulation other than a tendency to overly attribute significant sniff tests to bad luck.

**JEL classification:** C18, A14, B41

**Keywords:** publication bias, null hypothesis, specification test, balance test, placebo test

**Contact information:** Christopher Snyder, Department of Economics, Dartmouth College, 301 Rockefeller Hall, Hanover NH 03755, email: chris.snyder@dartmouth.edu. Ran Zhuo, Department of Economics, Harvard University, Littauer Center, 1805 Cambridge Street, Cambridge, MA 02138, email: rzhuo@g.harvard.edu.

**Acknowledgments:** The authors are grateful to the Sloan Foundation for financial support as well as to three programs at Dartmouth College (Dartmouth Economic Research Scholars, Lucas Family Fund for Undergraduate Research, and Presidential Scholars Program) for supplementary funding for some research-assistant time. The authors thank Barbara DeFelice, Ethan Lewis, Bruce Sacerdote, and Douglas Staiger for advice and suggestions. The authors are indebted to their dedicated team of research assistants: Kelsey Anspach, Ekshesh Bekele, Albert Chen, Barry Chen, Natalia Drozdoff, Isabel Hurley, Kirill Savolainen, and Olivia Zhao.

# 1. Introduction

Readers naturally focus on the significance of the main results in empirical articles. Our paper shifts the focus to the significance of auxiliary tests (including balance, specification, overidentification, falsification, and placebo tests) supporting the credibility of papers' main results. We dub such auxiliary tests "sniff tests" because of the subjective standards for passing and the bad news for the paper's main results that comes with rejection. Sniff tests provide a new and valuable lens through which to study publication bias in the academic literature.

Publication bias, formally defined as systematic error introduced by selective publication of evidence that is not representative of all available evidence (see, e.g., Bax and Moons 2011), has long been acknowledged as a threat to empirical inference. The large literature in the hard sciences as well as economics that has studied publication bias has focused on canonical forms such as the "file-drawer problem" (coined by Rosenthal 1979), whereby studies that find insignificant treatment effects remain in the metaphorical file drawer rather than being published. Results from published studies would thus constitute a sample selected on their significance, biasing the reported significance levels upward. Authors may also contribute to publication bias through deliberate actions to boost significance and thus publication chances (e.g., data mining or outright manipulation of data or results).

Sniff tests offer a unique lens through which to study publication bias. Authors have the reverse of the usual preferences with sniff tests, preferring them to be statistically insignificant. With standard tests of main treatment effects, published studies may only reveal a highly selected "tip of the iceberg" of the most significant results. With sniff tests, by contrast, though publication bias may chip away at the tip of the most significant results, it does not distort the bulk of insignificant results that otherwise lies below view with standard tests. Furthermore, forces behind publication bias may be more or less strong with sniff tests. An insignificant main result may be more likely to ruin a paper's chance of publication than a significant sniff test, leading to more publication bias with standard tests. On the other hand, it may be easier for authors to manipulate how sniff tests are reported. Editors always demand the reporting of main results but not necessarily every imaginable sniff test, allowing authors some leeway to suppress or bury significant sniff tests. Such considerations could potentially lead to greater publication bias with sniff tests.

There are a number of other advantages to studying publication bias through the lens of sniff tests. Perhaps the chief advantage is that the underlying distribution of probability values (p-values) from sniff tests is known for certain subsamples under certain conditions. Specifically, under the null hypothesis of no publication bias, p-values from tests of covariate balance between treatment versus control samples in a randomized controlled trial (RCT) should have a uniform distribution on $[0, 1]$ (as long as authors have not taken measures to improve balance). This property offers a powerful, reduced-form test of publication bias, comparing the observed distribution of a large number of test statistics against the theoretical distribution. Traditional methods to detect publication bias such as meta-regression analysis (see Stanley and Jarrell 1989, Stanley 2005, and Stanley 2008) operate on severely restricted samples—collections of studies that investigate the same pair of dependent and independent variables.

Our dataset consists of a painstakingly collected sample of nearly 30,000 sniff tests from 900 articles published in 60 economics journals. We first identified candidate articles via a search over relevant keywords. A team of research assistants scanned each of these candidate articles to verify whether or not it contained one or more tables of sniff-test results; and if so, p-values, level of significance, and other article and journal information was collected, down to details such as the position of the table in the paper, the symbols used to highlight significance, even the nature of author claims about how well the sniff test performed.

The analysis employs both reduced-form and structural methods. While we take an initial look at the the full dataset, most of the analysis focuses on what we call the "pure" sample of sniff tests, for which we are sure the authors did not take measures (re-randomization, stratification, or matching) to lessen the p-values' significance. We further break the pure sample down into balance tests in RCTs and other tests. The reason for analyzing these subsamples separately is that, as mentioned, the distribution of p-values is known to be uniform in the absence of publication bias for the pure sample of RCT balance tests. For other tests, the p-value distribution is unknown in the absence of publication bias. An unknown, perhaps substantial, number of studies may suffer flaws such as pre-trends or other misspecification that are detected by sniff tests, leading to more than, say, 10% of p-values being significant at the 10% level. Publication bias may reduce the excess of significant p-values but perhaps not enough to be detected.

Our reduced-form methods are straightforward, amounting to visual inspection of kernel den-

sities of various subsamples of p-values, and simple tests of the equality between the proportion of significant p-values and the uniform benchmark. For the pure sample of RCT balance tests, the mass missing from the $[0, 0.1)$ interval of p-values relative to the uniform case is consistent with nearly 45% of significant RCT balance tests remaining in the file drawer. For the pure sample of other tests, the kernel density of p-values looks quite different, with a huge spike of p-values near 0, consistent with a substantial number of studies suffering from flaws picked up by sniff tests.

For the subsample of tests other than of RCT balance, we employ structural methods to estimate the rate of publication bias. We specify a two-stage model in which a flexible beta distribution describes the rate of study flaws in the first stage. In the second stage, a proportion of studies with significant sniff tests, whether due to random chance or study flaws, are assumed to be removed by the publication process. Though simple, this model predicts a peculiar shape for the distribution of p-values, which turns out to fit the data quite well, lending credibility to the structural estimates. Our structural estimates imply that over 90% of significant tests in this subsample end up in the "file drawer." The overarching conclusion is that the publication process has a substantial, not an ignorable, effect on the distribution of published test statistics.

We are able to obtain some insight into potential mechanisms behind publication bias by looking at heterogeneity across categories. Reassuringly, we find no evidence author manipulation. Authors do not appear to hide significant sniff tests in appendices, obscure them by omitting significance symbols, nor engage in fudging or mining to move close p-values over crucial significance thresholds. We do find some evidence that sniff tests in top-tier journals are less significant than those in lower-tier journals, consistent with top journals having more rigorous standards for all aspects of empirical studies—laudable, though suggestive that publication bias may be particularly severe in the top tier. We also find that sniff tests are less significant in smaller tables and more recent articles.

Our estimate of the aggregate distribution of sniff-test p-values—to our knowledge the first in the literature—sheds further light on where individual sniff-test statistics stand in the empirical distribution and whether author claims associated with individual statistics are collectively justified. Given that standards for passing are subjective, to evaluate the performance of sniff tests in any individual paper, it is useful to know where those individual p-values fall in the broader distribution in the economics literature. If the aggregate distribution turns out to be uniform, this would lend

credence to claims that any individual sniff test that happens to be significant is a false positive rather than an actual study flaw. For example, if 10% of tests of pre-trends are significant at the 10% level, assuming there is no publication bias, one can conclude that many of these are due to random noise rather than a real trend in the data-generating process. Conversely, we can use the aggregate distribution to check author claims that their studies have no more significant sniff tests than would be expected at random. We end up rejecting such claims in the aggregate for the most significant threshold (1% level of significance), finding more than double proportion of p-values there than would be expected if such claims were true.

## 2. Literature Review

The large and growing literature on publication bias dates at least back to Sterling (1959); Rosenthal (1979); Begg and Berlin (1988); and, in the field of empirical economics, DeLong and Lang (1992). As discussed in reviews by Rothstein *et al*. (2006) and Christensen and Miguel (2016), the literature chiefly focuses on the bias arising from the predisposition of agents at all levels (authors, reviewers, readers) to favor statistically significant results in tests of null hypotheses that the main results of interest are zero. Statistically significant results are usually seen as more conclusive, leading authors to mine the data for significant specifications, leading authors to to be more likely to submit papers with significant results for publication, and leading journals to be more likely to publish submitted papers with significant results.[1] As modeled in Hedges (1992), this form of publication bias will shift reported p-values from high to low values.

To our knowledge, ours is the first paper to study bias arising from sniff tests, for which a statistically significant result is bad news for the study's methodology, leading agents to have the opposite predisposition, favoring insignificant results, providing a new window through which to study publication bias.

The largest body of evidence on publication bias is in the field of medicine. Besides being a large and high-stakes field, a further advantage is that medical trials often require prior registration. This allows meta-researchers to study the universe of all registered projects, both unpublished and

[1] Card and Krueger (1995) identify two other sources of publication bias relevant in economics if not other fields. First, reviewers may be predisposed to accept papers consistent with the conventional view; second, researchers may use the presence of conventionally expected results as a model-selection test. See the surveys by Stanley (2005) and Ioannidis and Doucouliagos (2013) for further discussion of publication bias in economics.

published, determining whether the distribution of test statistics differs across these two groups. Examples include Easterbrook *et al*. (1991); Dickersin, Min, and Meinert (1992); Stern and Simes (1997); Hopewell *et al.* (2009), and Driessen *et al.* (2015). Dickersin *et al.* (1987) tracked the progress of registered medical trials that failed to be published, finding that progress was typically halted by author discretion, with the authors either deciding not to write up the paper or not to submit a completed paper, rather than rejection of submitted manuscripts by the journals.

Opportunities to identify the universe of unpublished and published studies is rare for meta-researchers. More commonly, only the selected set of published articles can be observed. To facilitate the detection of publication bias in this selected set, meta-researchers have focused on isolated cases in which many studies of the same pair of dependent and independent variables have been published. Popular techniques used include the funnel plot (Egger, *et al*. 1997), rank correlation tests (Begg and Mazumdar 1994; Egger, *et al*. 1997; Sterne, Gavaghan, and Egger 2000; Duvall and Tweedie 2000; Peters *et al*. 2006; Stanley and Doucouliagos 2014), and parametric selection models (Iyengar and Greenhouse 1988, Hedges 1992, Andrews and Kasy 2017). Economists have applied these techniques to detect publication bias in a variety of applications (see Stanley 2005 for a survey). Examples include meta-analyses by Card and Krueger (1995) on the effect of minimum wage on employment; Ashenfelter, Harmon, and Oosterbeek (1999) on the effect of schooling on earnings; Görg and Strobl (2001) on the effect of multinationals on domestic productivity; Doucouliagos (2005) on economic freedom; Nelson (2014) on the price elasticity of beer; and Havranek (2015) on intertemporal substitution.

Commentators have proposed a variety of solutions to the publication-bias problem. The proposed methods include registering every trial, publishing every study (Thornton and Lee, 2000), and improving reporting protocols (Moher, *et al*. 2010). In economics, Duflo, Glennerster, and Kremer (2008) propose that granting agencies take the lead, requiring all studies they fund to be archived, even those that end up not being published. Some recent work explores the possibility of correcting for publication bias *ex post* (McCrary *et al*. 2016 and Andrews and Kasy 2017).

Besides making a novel contribution to the literature on publication bias, our paper provides what to our knowledge is the first meta-analysis of the important class of tests we have labeled sniff tests. Attempts have been made in the economic literature to study publication bias from the aggregate distributions of tests (see Brodeur, *et al*. 2016 and Ioannidis, Stanley, and Doucouliagos

2017). These articles focus on standard tests of main results, not sniff tests. As pointed out in the introduction, there are several advantages to studying publication bias through the lens of sniff tests. Perhaps most important is that the underlying theoretical distribution of certain sniff tests in the absence of publication bias is known, allowing powerful identification of publication bias using straightforward reduced-form methods. Closest in spirit to our paper is previous work by Bruhn and McKenzie (2009). In addition to a valuable survey of practice among leading authors and a Monte Carlo analysis, the authors review the results of balance tests in a collection of articles in development economics. While that study examined 13 articles, our sample includes nearly 900 articles across all fields of economics, allowing us to obtain a broader view of the distribution of p-values in the economics literature and to run formal statistical tests.

# 3. Data

We collected data on sniff tests by having a team of research assistants systematically examine an initial, large pool of journal articles in economics. We identified this initial pool of articles from ScienceDirect, Elsevier's online database of journal articles. We collected PDF files for all economics articles that were turned up by a search of related keywords such as "balance test," "baseline comparison," "falsification test," "placebo test," "randomization," "validation check," etc. We supplemented the Elsevier journals with five top-tier, general-interest journals in economics archived on JSTOR (*American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Review of Economic Studies*, *Quarterly Journal of Economics*), performing the same keyword search as on ScienceDirect. As the keywords were relatively uncommon before 2005, we restricted our pool to articles published from 2005 to 2015.[2]

A team of research assistants browsed each article in this initial pool, determining whether it contained a table reporting sniff tests. If so, the research assistant collected data on test statistics, p-values, and significance levels reported in the table or tables containing the sniff tests along with relevant table, article, and journal information. Only two-sided tests were included.[3] All work was

[2]JSTOR is missing the 2013–15 volumes of the *Quarterly Journal of Economics* and the 2015 volume of the *Review of Economic Studies*. We performed our standard keyword search on Google Scholar to identify the initial pool of articles in these volumes.

[3]We omitted one-sided tests given the difficulties of analyzing publication bias in the context of one-sided tests, outlined by DeLong and Lang (1992). However, we did notice some peculiar cases in our data collection in which

re-checked by supervising research assistants. We dropped 689 observations that either were not structured as well-defined hypothesis tests with associated p-values or did not provide sufficient information to glean an exact p-value or an interval for it. Dropping a further observation with a p-value exceeding 1 leaves a final dataset of 29,812 sniff-test observations. These observations are reported in 1,391 different tables from 890 articles published in 59 journals. Appendix Table A1 lists the journals and the contribution each makes to the sample. The *Journal of Health Economics* contributes the most observations (16% of the sample), followed by the *Journal of Public Economics* (11%), the *Journal of Development Economics* (10%), *Labour Economics* (8%), and the *American Economic Review* (8%).

Figure 1 graphs the number of observations in our dataset by year of publication of the containing article. To the extent the journals in our sample are representative, this figure provides a picture of the growing use of sniff tests in the economics literature. Starting from few sniff tests in 2005, the number of sniff tests in our dataset grows at a 40% average annual rate.

Table 1 presents subsample breakdowns for the dataset. The first breakdown is by methodology. This is a key breakdown: it will be important to distinguish among methodologies because different types of tests can have very different distributions of p-values. Balance tests from randomized controlled trials (RCTs) will have a special place in our analysis because their p-values have a known distribution under the null hypothesis of no publication bias, in particular, the uniform $[0,1]$ distribution. Other tests (balance tests from non-RCTs, placebo tests, and various types of falsification and specification tests) do not necessarily entail a uniform distribution in the absence of publication bias. Instead, the distribution will depend on the nature (unknown to us econometricians) of the underlying data and model appearing in papers in our dataset. For example, specification tests will exhibit a concentration of low p-values if many papers in our dataset hap-

authors used an inconsistent reporting convention, gauging significance levels for their sniff test according to the tight thresholds of a two-sided test but then discussing the test as if it were one-sided. For example, the authors might test for a pre-trend, discover its p-value is significant at the 5% level according to a two-sided test, but then argue that the trend is in the opposite direction from their result of interest (say a downward trend working against a positive jump at a regression discontinuity). If in fact the test cannot be rejected when a negative trend is found, it was a mistake to have used two-sided rather than one-sided significance thresholds. Of course the one-sided null would not be rejected in this particular test regardless of significance level used because the trend was on the "accept" side of the inequality. The problem raised by this selective use of one- and two-sided tests is in the interpretation of other sniff tests in the article: one-sided tests may have been appropriate for these, too, but rejections glossed over by the use of tighter two-sided thresholds. It is unclear how best to treat these peculiar cases of one-sided tests posing as two-sided. We checked the robustness of all of our results including and excluding them. The results are very similar in magnitude and significance. All results presented in this paper include these observations.

pen to suffer from misspecification. On the basis of these considerations, some of our analysis will focus exclusively on balance tests in RCTs, which constitute 7,150 observations, or 24% of the sample; analysis of other types of tests in the remaining 76% of the sample will require suitably careful interpretation.

Valid techniques exist to improve balance in RCTs relative to what we will label the "pure" case of a single randomization. These techniques including re-randomization (randomizing treatment/control selection multiple times until a desired balance in covariates between groups is achieved), stratification (randomizing treatment/control selection within covariate strata), and matching (finding pairs of observations with the same covariates, making one a treatment and the other a control).[4] Tautologically, application of balance-improvement techniques raises the p-values of the subsequent balance tests. If we included balance tests that have be run after balance improvement in our sample of RCT balance tests, the distribution of p-values would no longer be expected to be uniform in the absence of publication bias. Instead, some of the mass of low p-values would be shifted to higher values, creating a false impression of the existence of publication bias. To avoid this problem, we construct a pure subsample, separating the remaining items for which there had been some balance improvement into different subsamples.[5]

Bruhn and McKenzie's (2009) survey of leading researchers who conduct RCTs in developing countries reveals a hodgepodge of approaches toward balance. While it is not uncommon for researchers to resort to a single randomization (80% did so for at least one past experiment; 40% did so for their most recent experiment), balance-improvement techniques are also widely used. Their review of selected publications suggests that when authors used re-randomization, the process was not described in detail. To account for opaque or even possibly missing mentions of re-randomization, we are conservative in our formation of the pure subsample, only include items in it when we can rule out re-randomization out on *a priori* grounds. Among other exam-

[4] For a recent, general discussion of these methods, see Athey and Imbens (2017).

[5] A potential concern is that the analysis of a subsample selected on the employment of balance-improvement techniques could lead to an endogeneity bias. Supposing that better researchers are more likely to employ balance-improvement techniques, the pure subsample would over-represent the work of less-skilled researchers. Supposing, in turn, that researcher skill is correlated with the performance of the balance test—positively correlated if less-skilled researchers botch the randomization procedure more often, negatively correlated if less-skilled researchers face stricter standards on balance tests to offset problems elsewhere in the paper—this could bias the pure subsample toward more or less significant balance tests. As discussed in the next paragraph, we select the pure sample based on whether or not re-randomization was possible rather than whether or not it was actually undertaken, alleviating this endogeneity problem.

ples, this includes cases in which a public lottery determined treatment status and cases in which treatment had been assigned by a third-party who was not concerned with statistical analysis. Our pure subsample contains 42% of RCT balance test items. For 14% of RCT balance tests, though re-randomization is not mentioned in the article, we cannot definitively rule it out because authors had access to the baseline data to re-randomize before treatment assignment. We classify this subsample as "possibly pure." There is less concern about the opacity of the reporting of stratification and matching; perhaps owing to the deliberation involved in their application, they are thought to be reported whenever used. In sum, therefore, observations in our possibly pure sample are definitively not stratified or matched; no mention is made about re-randomization although we cannot definitively rule it out. For the remaining 44% of RCT balance tests, one or more of re-randomization, stratification, or matching was definitively used to improve balance.[6]

Turning to the subsample of tests other than balance in RCTs, re-randomization and stratification are irrelevant in this context since authors do not have the required control over randomization required by these techniques. Matching is the remaining, feasible method for p-value improvement, which authors employ by restricting analysis to a matched subsample of their full sample. In 76% of these other tests, no matching, and thus no p-value improvement method, was employed. The rest involve some matching. For 8% of other tests, matching is employed and results of sniff tests prior to matching are reported. For 16% of other tests, matching is likewise employed but results of sniff tests after matching are reported.

For the test statistics collected, we tried to glean p-values whenever possible, whether directly reported by the authors or whether the authors supplied enough ancillary information for us to compute the p-value. We were able to glean exact p-values for 41% of the observations. For 59% of the observations, we were unable to glean an exact p-value but the authors did report the interval in which the p-value fell, usually indicated by asterisks alongside the reported test statistic.

We break down our sample in several additional ways to be used for more nuanced tests of the "file-drawer problem" and of manipulation. First, we compare test statistics reported in different

[6] Stratification need not compromise the uniform distribution of p-values under the null hypothesis of no publication bias. Of course stratification improves treatment/control balance for the stratifying variables by construction, but these variables are seldom included in subsequent balance tests. Instead, balance in other variables is tested. If these other variables are uncorrelated with the stratifying variables, then balance-test p-values can still be uniform under the null of no publication bias. On the other hand, if these other variables are correlated with the stratifying variables, balance in the stratifying variables may be inherited by correlated variables, reducing the mass of significant p-values from balance tests compared to a uniform distribution.

locations in an article, body versus appendix. If reviewers are less likely to notice problematic p-values reported in appendices, we may observe higher p-values in the body subsample (72% of observations) than the appendix sample (28% of observations). The effect will be exacerbated if the placement of results is endogenous, with authors more likely to shift problematic results into the appendix. Second, we compare test statistics by the authors' reporting convention. Using symbols to highlight significant tests (65% of observations used such symbols either alone or together with a report of the exact p-value) may draw more attention to unfavorable sniff tests than when symbols are omitted (35% of observations). This may lead to a correlation between the reporting convention and p-values if publication bias or manipulation is present. Third, we compare test statistics reported in different tiers of journals. The stringent requirements for publication at top-tier may include less tolerance for low p-values. If so, we may see a systematic difference in p-values across journal tiers, with lower p-values in lower-tier journals. We divided journals into tiers on the basis of five-year impact factors from *Thomson Reuters Journal Citation Reports 2015,* with tier-1 journals having a five-year impact factor above four, tier-2 journals between two and four, and tier-3 journals below two. See Appendix Table A1 for the list of journals' tiers and impact factors. The resulting ranking is sensible, with tier 1 including the five general-interest journals archived on JSTOR as well as the top Elsevier journals in the subfields of finance and accounting (respectively, the *Journal of Financial Economics* and *Journal of Accounting and Economics*). Roughly a quarter of the observations are in the top tier, a quarter in the bottom tier, and half in the middle tier.

We also collected information on authors' stated satisfaction with the test outcome. Our research assistants read the text passage in each article discussing the relevant test statistics and rated the strength of authors claims according to four categories, "strong claim," "weak claim," "no claim," and "admit rejected," based on the following rubric. The "strong claim" category includes cases in which the authors express satisfaction with the test outcome. Authors express satisfaction—with good reason—when the associated p-value under consideration is insignificant. We also consider authors to be strongly satisfied whenever, faced with having to explain a significant p-value, they explicitly attribute it to random chance rather than some systematic feature of the data. These cases are often associated with tables reporting multiple sniff tests, only a few of which are significant. The "weak claim" category includes cases in which, faced with some

significant sniff-test results to explain, perhaps too many to attribute to random chance, the authors are forced to acknowledge possible problems at the same time mounting some defense of their results. Typical of this category is for authors to acknowledge that the test outcome indicates the existence of imbalance or pre-treatment effects but then argue that it does not undermine the validity of their main results, often using the argument that the significant sniff-test results follow no systematic patterns. When the authors do not discuss the specific test statistic, we classify it as "no claim." "Failed claim" includes cases when authors freely acknowledge that the significant p-value indicates rejection of the sniff test and a potential problem for their study. The majority of observations (71%) are classified as "strong claims," reflecting in part the frequency of tables lacking problematic p-values to begin with. "Weak claims" account for 9% of observations, "no claims" for 7%, and "failed claim" for 13% of observations.

# 4. Theory

This section provides the theoretical underpinnings for our empirical analysis in a series of subsections. The first subsection derives the distribution of our main object of study, the p-value, under the null hypothesis of no publication bias or study flaws. The second subsection explores which among alternative conditional expectations is the most natural measure of a departure of the population of p-values from the null. The analysis in this subsection will help inform the appropriate reduced-form estimator. The last subsection constructs a structural model illustrating possible shapes of the p-value distribution when the null is violated.

## 4.1. Null Distribution of P-Values

Let $i \in I$ index individual sniff-test observations, where $I$ denotes the population of sniff tests. The key variable in our analysis is the p-value, denoted $p_i$, associated with sniff test $i$. To define $p_i$ formally requires some notation. A sniff test is the test of the null hypothesis $H_0$ of no study flaw against the alternative $H_1$ of a flaw of whatever sort under consideration. The author computes the relevant test statistic, denoted $Z_i$, from the data and sees whether it lies in rejection region $\Gamma(\alpha)$, depending on a pre-specified significance level, $\alpha$. For one-sided tests, $\Gamma(\alpha)$ typically includes all test statistics above a threshold; for two-sided tests, $\Gamma(\alpha)$ typically includes all test statistics whose

absolute value is above a threshold. If $Z_i \in \Gamma(\alpha)$, then $H_0$ is rejected in favor of $H_1$ at the $\alpha$ level; otherwise $H_0$ cannot be rejected.

Assume the test statistic is continuously distributed. Under this assumption, the definition of significance level entails

$$\Pr(Z_0 \in \Gamma(\alpha)) = \alpha, \tag{1}$$

where $Z_0$ is a draw of the test statistic under the null hypothesis. We will further require that to be well-defined, $\Gamma(\alpha)$ must be an increasing correspondence; i.e., for $\alpha', \alpha'' \in [0,1]$,

$$\alpha' < \alpha'' \iff \Gamma(\alpha') \subset \Gamma(\alpha''). \tag{2}$$

Define the significance $A(Z_i)$ of a test statistic $Z_i$ as the significance of the smallest rejection region containing $Z_i$:

$$A(Z_i) = \underset{\alpha \in [0,1]}{\operatorname{argmin}} \{\Gamma(\alpha) | Z_i \in \Gamma(\alpha)\}. \tag{3}$$

The p-value can then be defined as

$$p_i = \Pr(Z_0 \in \Gamma(A(Z_i))), \tag{4}$$

i.e., the probability that the test statistic computed under the null is at least as significant as that computed from the data.

The "textbook" result that p-values have a uniform $[0,1]$ distribution under the null turns out to hold in our model in the absence of publication bias. Appendix B provides a formal proof. This result will lead us to use the uniform distribution as a natural benchmark in our subsequent analysis. Unfortunately for authors in our data, the null is not guaranteed to hold in their studies; some studies will exhibit failures of balance, misspecification, pre-trends, and so forth. In addition, a key issue of interest to us, publication bias, will alter the distribution of p-values from the benchmark. The next subsection discusses some econometric issues associated with reduced-form estimation of p-values' significance.

## 4.2. Reduced-Form Model

Our estimates of the proportion of significant p-values in the population will depend on our sampling frame. In this subsection, we provide a reduced-form model allowing us to determine which sampling frame is most natural and characterize the biases arising from other sampling frames. Besides clarifying the discussion, the model directly leads to the reduced-form estimator proposed in Section 5.

The broad motive for our study is to detect whether the typical piece of scholarship exhibits publication bias and/or misspecification. This naturally suggests a sampling frame in which scholarly works—we take tables as our unit of work—are collected and the average proportion of significant p-values across them is compared to the null. An alternative sampling frame collects all the sniff tests together and randomly sample from the collection. We will show that if different scholarly works contain different numbers of sniff tests and the number of sniff tests presented in the work is correlated with their significance, then the two sampling frames can generate different expectations. It is thus important to choose the appropriate sampling frame and focus the analysis on the expectation associated with that frame.

To formalize these ideas, let $t \in T$ index tables, which we take as the unit of scholarly work, where $T$ denotes the population of tables. We sometimes emphasize the "containing" relation with functional notation, letting $t(i)$ denote the table containing sniff test $i$. Let $n_t$ denote the number of sniff tests in table $t$.

The precision at which p-values are reported differs across tables, some specifying an interval for $p_i$, others the exact value. To formalize these reporting conventions, let $R_t$ denote the significance thresholds reported by table $t$. For example, if table $t$ does not directly report p-values, only indicating significance at the 5% level with a star, then $R_t = \{0.05\}$. If table $t$ again does not directly report p-values but indicates significance at the 10%, 5%, or 1% levels with one, two, or three stars, then $R_t = \{0.01, 0.05, 0.1\}$. If table $t$ directly reports p-values, then, abstracting from rounding issues, $R_t = (0,1)$. Let $\alpha \in (0,1)$ denote the significance threshold that we want to study. A table's reporting convention may or may not line up with $\alpha$. Let $T_\alpha = \{t \in T \,|\, \alpha \in R_t\}$ denote the subpopulation of tables reporting significance at the $\alpha$ level and $I_\alpha = \{i \in I \,|\, \alpha \in R_{t(i)}\}$ denote the subpopulation of sniff tests in that subpopulation of tables.

Let $S_{i\alpha} = \mathbf{1}(p_i < \alpha)$ be the indicator of whether sniff test $i$ is significant at the $\alpha$ level. We are

interested in the proportion of significant sniff tests in the typical table, captured by the following conditional expectation:

$$\pi_\alpha = E_{t \in T_\alpha}(E(S_{i\alpha} \mid i \in t)). \tag{5}$$

This corresponds to the sampling frame that first samples tables and then sniff tests within tables. An obvious alternative is to directly sample from the population of sniff tests, leading to the expectation $E_{i \in I_\alpha}(S_{i\alpha})$. At first glance, this seems simpler than (5), but after rearranging into a form more comparable to (5), we see it involves an extra factor:

$$E_{i \in I_\alpha}(S_{i\alpha}) = E_{t \in T_\alpha}\left(\frac{n_t}{\bar{n}_\alpha} E(S_{i\alpha} \mid i \in t)\right), \tag{6}$$

where $\bar{n}_\alpha = E_{t \in T_\alpha}(n_t)$ denotes expected table size.

It is immediate that if tables are all the same size, then (5) equals (6) because $n_t = \bar{n}_\alpha$. To provide a more general comparison of the two expectations, we introduce the following reduced-form model. By equation (5), we can write $S_{i\alpha}$ as

$$S_{i\alpha} = \pi_\alpha + u_{i\alpha}, \tag{7}$$

where $u_{i\alpha}$ is an error term with conditional expectation

$$E_{t \in T_\alpha}(E(u_{i\alpha} \mid i \in t)) = 0. \tag{8}$$

Assume the error term can be decomposed as $u_{i\alpha} = \nu_{t(i)\alpha} + \varepsilon_{i\alpha}$, where $\varepsilon_{i\alpha}$ is an idiosyncratic error having mean zero within and across tables, i.e.,

$$E(\varepsilon_{i\alpha} \mid i \in t) = 0, \tag{9}$$

and where $\nu_{t(i)\alpha}$ is unobserved table effect. By (8) and (9), $\nu_{t\alpha}$ has mean zero across tables, but we allow $\nu_{t\alpha} \neq 0$ within any given table.

Substituting this series of assumptions into (6) and rearranging yields

$$E_{i \in I_\alpha}(S_{i\alpha}) = \frac{1}{\bar{n}_\alpha} E_{t \in T_\alpha}(E(n_t\pi_\alpha + n_t\nu_{t(i)\alpha} + n_t\varepsilon_{i\alpha} \mid i \in t)) = \pi_\alpha + \frac{1}{\bar{n}_\alpha}\,\mathrm{Cov}_{t \in T_\alpha}(n_t, \nu_{t\alpha}). \tag{10}$$

We see that (6) is biased relative to $\pi_\alpha$—positively biased if $v_{t\alpha}$ covaries positively with table size $n_t$ and negatively biased if $v_{t\alpha}$ covaries negatively with $n_t$.

An additional nuance in the expectation (5) is that, rather than being taken over the entire population of tables $T$, the expectation is restricted to the subpopulation $T_\alpha$. This is done for two reasons. First, it is sometimes impossible to compute $S_{i\alpha}$ for $t(i) \in T \setminus T_\alpha$. For example, in studying the threshold $\alpha = 0.05$, suppose we have a starred observation $i$ from a table that only reports significance at the 10% level; formally, $S_{i,0.10} = 1$ and $R_{t(i)} = \{0.1\}$. We know $p_i \in [0, 0.1)$ for that observation but not whether $p_i \in [0, 0.05)$ or $p_i \in [0.05, 0.10)$. Hence, $S_{i\alpha}$ cannot be computed in this example. A second reason for conditioning on $T_\alpha$ rather than $T$ can be understood by returning to the previous example but modifying it so that $i$ is now an unstarred entry in its containing table. It is possible to compute $S_{i\alpha}$ in this modified example: insignificance at the 10% level implies insignificance at the 5% level, so $S_{i,0.1} = 0$ implies $S_{i,0.05} = 0$. However, such tables only supply insignificant observations, biasing the estimate of $\pi_\alpha$ downward. The bias can go the other way as demonstrated in an example in which we are studying the threshold $\alpha = 0.05$ and have a table reporting significance at the 1% level; $S_{i\alpha}$ can only be computed from starred sniff tests from that table, leading to an upward selection bias. Restricting the population to $T_\alpha$ eliminates these biases.

## 4.3. Structural Model

In this subsection, we construct a structural model that illustrates what the distribution of p-values looks like for different departures from the null hypothesis.

P-values are generated by a two-stage process. In the first stage, p-values are drawn from the population of studies exhibiting a certain rate of flaws. We will adopt a specific functional form both to fix ideas in the present discussion and to form the basis of the structural estimation in Section 6. Building on Sellke, Bayarri, and Berger's (2001) suggestion of using a beta distribution with one free parameter as the alternative to a uniform distribution of p-values, assume that the initial draw of p-values follows a the beta distribution with parameters $1/(1+m)$ and 1, having the associated density

$$f_1(p_i; m) = \left[(1+m)p_i^{\frac{m}{1+m}}\right]^{-1}. \tag{11}$$

Parameter $m \in [0, \infty)$ captures the extent of flaws/mistakes in the population of studies. When

$m = 0$, there are no flaws in the population of studies, and $f_1$ reduces to the uniform density. On the other hand, when $m$ grows large, $f_1$ piles up all the mass of the distribution on the lowest p-values.

In the second stage, the file-drawer problem or other feature of the publication process results in a proportion $b \in [0,1]$ of the p-values significant at better than some cutoff level $\hat{\alpha}$ being removed from the published sample. Applying Bayes Rule to the prior density in equation (11), the posterior density emerging from the second stage, the density characterizing p-values observed in our sample, is

$$f(p_i; m, b, \hat{\alpha}) = \frac{1 - S_{i\hat{\alpha}} b}{1 - \hat{\alpha}^{\frac{1}{1+m}} b} f_1(p_i; m). \tag{12}$$

The assumption that parameters $b$ and $m$ are common across observations simplifies the analysis and the later structural estimation; but note this imposes a restriction on the reduced-form model from the previous subsection, which allowed tables to have different values of $\nu_{t\alpha}$. (We could adapt the structural model to allow for heterogeneity across tables by subscripting $b_t$ and $m_t$ and specifying a joint distribution of these parameters across the population of tables $T$.)

Figure 2 illustrates the shape of this density for various parameters. The top panel illustrates the density in the absence of publication bias but in the presence of study flaws, i.e., $b = 0$ but $m > 0$. Note this just the first-stage density, as $f(p_i; m, 0, \hat{\alpha}) = f_1(p_i; m)$. These densities have disproportionately greater mass on low p-values than the uniform distribution, increasingly more mass the higher is $m$.

The middle panel illustrates the density when there is publication bias but no study flaws, i.e., $b > 0$ but $m = 0$. The cutoff for potential removal is set to $\hat{\alpha} = 0.1$. These densities have disproportionately less mass on p-values below $\hat{\alpha}$ than would the uniform distribution, with less mass there the higher is $b$. Substantial publication bias can result in a quite substantial reduction in the mass of p-values below $\hat{\alpha}$.

The bottom panel illustrates the case with both publication bias and study flaws, i.e., $b, m > 0$. In this case, the mass of p-values in the below $\hat{\alpha}$ may be greater or less than the uniform benchmark since the two forces have opposing effects on this mass. For any degree of publication bias $b$ and any removal cutoff $\hat{\alpha}$, one can compute the degree of study flaws $M(b, \hat{\alpha})$ that results in the two forces exactly offsetting, so that the mass of p-values below $\hat{\alpha}$ equals the mass $\hat{\alpha}$ under a uniform

distribution:

$$M(b,\hat{\alpha}) = \frac{\ln(1-(1-\hat{\alpha})b))}{\ln\hat{\alpha} - \ln(1-(1-\hat{\alpha})b))}. \tag{13}$$

In the bottom panel of Figure 2, the exact balance between publication bias and study flaws has been achieved by setting $m = M(b,\hat{\alpha})$ for $\hat{\alpha} = 0.1$ and for various values of $b$. If $m > M(b,\hat{\alpha})$ in a given sample, then the mass of p-values will exceed the uniform benchmark $\hat{\alpha}$; if $m < M(b,\hat{\alpha})$, then the mass of p-values will be less than $\hat{\alpha}$.

This discussion of the figure highlights some key issues in identifying publication bias in our data. Finding that the mass of p-values in the interval $[0,\alpha)$ exceeds $\alpha$ is not necessarily evidence against publication bias; it could be that there is substantial publication bias but it is masked by the prevalence of study flaws. On the other hand, finding that the mass of p-values in the interval $[0,\alpha)$ is significantly less than $\alpha$ *is* evidence of publication bias. However, this may be a low-power test of publication bias because it would have to be extensive enough to overcome study flaws in the population that would tend to produce the opposite result. We isolate a subsample that resemble the middle panel of Figure 2, where we think $m \approx 0$, so that publication bias is not masked by an opposing effect. This will motivate our focus on the pure sample of balance tests in RCTs.

# 5. Reduced-Form Analysis

For most of our analysis, we are able to employ straightforward reduced-form methods, to which we turn next. Section 5.1 describes the reduced-form methods used. Sections 5.2–5.7 present the results for various subsamples and specifications.

## 5.1. Methodology

We previously argued that $\pi_\alpha$ in equation (5) is the appropriate parameter to estimate to judge whether sniff tests are disproportionately significant at the $\alpha$ level. This section presents the methods used to consistently estimate $\pi_\alpha$. We carry all the notation over from that earlier section with the exception that $T$ now represents the sample of tables and $I$ the sample of sniff tests rather than the respective populations. Similarly, we redefine $T_\alpha$ and $I_\alpha$ so that they are the subsamples rather than subpopulations as they were formerly defined. Vertical bars denote the cardinality of sets, so for example $|T|$ denotes the number of tables and $|I|$ the number of sniff tests in the sample.

By the weak law of large numbers, a consistent estimator of $\pi_\alpha$ can be obtained by replacing the expectations in (5) by sample averages:

$$\hat{\pi}_\alpha = \frac{1}{|T_\alpha|} \sum_{t \in T_\alpha} \left( \frac{1}{n_t} \sum_{i \in t} S_{i\alpha} \right). \tag{14}$$

An equivalent expression for $\hat{\pi}_\alpha$ that is particularly convenient can be obtained by introducing inverse frequency weights

$$w_i = \frac{|I_\alpha|}{|T_\alpha| n_{t(i)}}. \tag{15}$$

These weights are inversely proportional to the number of sniff tests in the including table $n_{t(i)}$, scaled by the constant necessary to have the weights sum to $|I_\alpha|$, the number of relevant observations. The necessary scaling constant equals $|I_\alpha| / |T_\alpha|$, as can be verified by summing (15):

$$\sum_{i \in I_\alpha} w_i = \frac{|I_\alpha|}{|T_\alpha|} \sum_{t \in T_\alpha} \left( \sum_{i \in t} \frac{1}{n_t} \right) = \frac{|I_\alpha|}{|T_\alpha|} \sum_{t \in T_\alpha} 1 = |I_\alpha|. \tag{16}$$

Using (14) and (15), we have

$$\hat{\pi}_\alpha = \frac{1}{|I_\alpha|} \sum_{i \in I_\alpha} w_i S_{i\alpha}, \tag{17}$$

the inverse-frequency-weighted average of $S_{i\alpha}$ for the subsample of sniff tests in tables directly reporting significance threshold $\alpha$.

The estimator in (17) involves two departures from the simple average of $S_{i\alpha}$. The inclusion of inverse frequency weights $w_i$ circumvents the potential bias that would arise from oversampling results from large tables that turn out to contain more or less significant results than average. The unweighted average, which does not correct for such oversampling, *is* an unbiased estimate of a certain conditional expectation, namely (6), but not the conditional expectation of interest here, $\pi_\alpha$.[7]

The second departure of (17) from the simple average of $S_{i\alpha}$ is that only observations in $I_\alpha$ are included; observations in tables not directly reporting significance at the level $\alpha$ under study are

[7] See Wooldridge (2010), chapter 20, for a textbook discussion of the need for weighting in various sampling contexts. Equation (14) is an implementation of the estimator he suggests for cluster-sampling contexts (see his equation (20.48)). Equation (17) is an implementation of the estimator he suggests for the standard-stratified-sampling context (see his equation (20.13)). In our application, the two contexts are equivalent since, within each stratum (i.e., each table), all observations (i.e., all sniff tests) are collected.

excluded. The selection bias that can arise when observations in $I \setminus I_\alpha$ are included was discussed in Section 4.2. The subsequent analysis will focus on the three most widely used significance thresholds in the literature: 1%, 5%, and 10%. Let $R^* = \{0.01, 0.05, 0.10\}$ denote the set of these three thresholds. If the intersection between $R^*$ and $R_t$ (the thresholds directly reported in table $t$) is empty, then all the observations in table $t$ are dropped from our analysis. Of the full sample of 29,812 observations described in Table 1, 101 do not report any threshold in $R^*$. Dropping them results in a final sample of size 29,711 used in the subsequent analysis.

Regression-based methods can be used to recover $\hat{\pi}_\alpha$. In particular, in the inverse-frequency-weighted least squares (IFWLS) regression of $S_{i\alpha}$ on a constant using the subsample $i \in I_\alpha$, the coefficient on the constant term is is numerically identical to $\hat{\pi}_\alpha$ in equation (14):

$$\hat{S}_{i\alpha} = \hat{\pi}_\alpha \cdot 1 \qquad i \in I_\alpha. \tag{18}$$

We use the notational convention of a hat on a dependent variable such as $\hat{S}_{i\alpha}$ to denote its fitted value from a regression, allowing us to distinguish estimating equations such as (18) from corresponding population models such as (7). In practice we will used the regression-based method to compute $\hat{\pi}_\alpha$ because this method conveniently generates clustered standard errors for hypothesis testing and readily extends to more complex models considered below.[8]

Rather than running separate regressions for each $\alpha \in R^*$, the estimates can be conveniently recovered from a single regression in which multiple copies of each observation $i$ are stacked, one copy for each threshold $\alpha \in R^* \cap R_{t(i)}$ that constitutes a match between our study set and the containing table's reporting convention:

$$\hat{S}_{i\alpha} = \sum_{r \in R^*} \hat{\pi}_r \mathbf{1}(\alpha = r) \qquad \alpha \in R^*, i \in I_\alpha. \tag{19}$$

The estimates $\hat{\pi}_\alpha$ in (18) and (19) are numerically equal, both equal to the proportion in (17). The clustered standard errors are also identical across equations (18) and (19). We cluster all standard errors at the table ($t$) level; this clustering strategy correctly adjusts for the appearance of multiple copies of each observation $i$ in the stacked regression.

[8] IFWLS regressions can be run in Stata using the `reg` command, setting analytic weights (`aweights`) equal to $1/n_{t(i)}$. Stata's automatic scaling using this command generates weights $w_i$.

The null typically used for hypothesis testing, $\pi_\alpha = 0$, is of little interest here. Under this null, there are no p-values in $[0, \alpha)$, more consistent with extreme publication bias than its absence. A more natural null hypothesis for this analysis is $\pi_\alpha = \alpha$, corresponding the mass of p-values that would arise under a uniform $[0,1]$ distribution, the distribution of p-values in the absence of misspecification and publication bias. We will conduct separate tests of $\pi_\alpha = \alpha$ for each $\alpha \in R^*$ as well as the joint test for all $\alpha \in R^*$.

The joint test provides a qualitative measure of the overall departure from uniformity across thresholds $\alpha \in R^*$. To proceed to quantify the common departure from uniformity, suppose first that the p-value proportions exceed the uniform thresholds by a constant level $\lambda$: i.e., $\pi_\alpha = \alpha + \lambda$ for all $\alpha \in R^*$. Substituting the analogous equation involving estimators into (19) and rearranging yields $\hat{S}_{i\alpha} = (\alpha + \hat{\lambda}) \sum_{r \in R^*} \mathbf{1}(\alpha = r) = \alpha + \hat{\lambda}$. Defining $Y_{i\alpha} = S_{i\alpha} - \alpha$, the significance indicator normalized by its expected threshold value under a uniform distribution, and $\hat{Y}_{i\alpha} = \hat{S}_{i\alpha}$ to be its fitted value from a regression, the preceding calculations yield the following regression specification:

$$\hat{Y}_{i\alpha} = \hat{\lambda} \cdot 1 \qquad \alpha \in R^*, i \in I_\alpha. \tag{20}$$

According to this equation, an estimate of the common level departure from uniformity $\hat{\lambda}$ can be recovered from an IFWLS regression of the normalized significance indicator $Y_{i\alpha}$ on a constant; the coefficient on the constant term provides the desired estimate, $\hat{\lambda}$. Suppose instead that the p-value proportions exceed the uniform thresholds by a constant proportion $\rho$: i.e., $\pi_\alpha = (1+\rho)\alpha$ for all $\alpha \in R^*$. Substituting the analogous equation involving estimators into (19) and rearranging yields $\hat{S}_{i\alpha} = (1+\hat{\rho})\alpha \sum_{r \in R^*} \mathbf{1}(\alpha = r) = (1+\hat{\rho})\alpha$, or upon further rearranging,

$$\hat{Y}_{i\alpha} = \hat{\rho}\alpha \qquad \alpha \in R^*, i \in I_\alpha. \tag{21}$$

According to this equation, an estimate of the common proportional departure from uniformity $\hat{\rho}$ can be recovered from regressing $\hat{Y}_{i\alpha}$ on $\alpha$ excluding a constant; the coefficient on $\alpha$ provides the desired estimate, $\hat{\rho}$. Agnostic about whether the departure from uniformity is in levels or proportions, we will estimate the departure both ways, as $\hat{\lambda}$ and $\hat{\rho}$. The qualitative results turn out to be similar either way.

It is straightforward to extend the methodology to allow for heterogeneity in $\lambda$ or $\rho$ across

categories of observations and estimate differences across the categories. For simplicity, suppose observations can be partitioned into two categories, say according to whether the observation appears in the body or appendix or whether its significance is highlighted by stars or not; the analysis readily extends to partitions with more than two categories. Let $C_i^k$ denote the indicator for whether $i$ belongs to category $k$. Suppose first that both categories depart from uniformity by different levels. Letting $\lambda^k$ denote the level departure from uniformity for category $k$, its estimate by $\hat{\lambda}^k$, and the difference between the estimates by $\Delta\hat{\lambda} = \hat{\lambda}^2 - \hat{\lambda}^1$, the following regression can be used to recover the estimates:

$$\hat{Y}_{i\alpha} = \hat{\lambda}^1 + \Delta\hat{\lambda} C_i^2 \qquad \alpha \in R^*, i \in \Omega_\alpha. \tag{22}$$

This equation implies that a consistent estimator of $\Delta\hat{\lambda}$ can be obtained by simply adding an indicator for the second group, $C_i^2$, to the stacked regression with a constant term, (20); the coefficient on the added indicator provides the desired estimate, $\Delta\hat{\lambda}$.

Suppose instead that both categories depart from uniformity by different proportions. Letting $\rho^k$ denote the proportional departure from uniformity for category $k$, its estimate by $\hat{\rho}^k$, and the difference between the estimates by $\Delta\hat{\rho} = \hat{\rho}^2 - \hat{\rho}^1$, the following regression can be used to recover the estimates:

$$\hat{Y}_{i\alpha} = \hat{\rho}^1 \alpha + \Delta\hat{\rho}\, \alpha C_i^2 \qquad \alpha \in R^*, i \in \Omega_\alpha. \tag{23}$$

This equation implies that a consistent estimator of $\Delta\hat{\lambda}$ can be obtained by simply adding the interaction between $\alpha$ and $C_i^2$ to the stacked regression on $\alpha$ alone in (21); the coefficient on the added interaction term provides the desired estimate, $\Delta\hat{\rho}$. Again, since we are agnostic about whether the difference between categories should be a level or proportional effect, we will estimate the difference both ways, as $\Delta\hat{\lambda}$ and $\Delta\hat{\rho}$. The qualitative results again turn out to be similar either way. Whether the difference across categories is assumed to be in levels or proportions, the appropriate null for testing these differences is the usual one: $\Delta\lambda = 0$ or $\Delta\rho = 0$, .

At times, we will be interested in visualizing the whole distribution of p-values rather than looking at the mass in a particular interval. To do so, we will plot histograms or, when smoothing aids visualization, kernel density estimates.[9] Both histograms and kernel density estimates are

[9]While the typical application of kernel density estimation is to random variables with unbounded supports, in our application the support of the distribution of p-values is $[0,1]$, having both lower and upper bounds. The kernel density estimator must be modified to account for these bounds. We perform the kernel density estimation using

based on the subsample for which we have exact p-values rather than intervals.

## 5.2. Results for Combined Sample

Table 2 presents estimates $\hat{\pi}_\alpha$ of the proportion of p-values significant at various thresholds. The estimates are computed using the stacked IFWLS regression (19). Standard error clustered at the table ($t$) level are reported in parentheses below the estimates. The size of the underlying subsample as well as the number of stacked observations and clusters are also reported.

Column (1) presents aggregate results for the full sample. The first entry shows that 6.1% of tests are significant at the 1% level, 9.7% are significant at the 5% level, and 12.8% are significant at the 10% level. If p-values were uniformly distributed on $[0, 1]$, we would expect the percent to match the significance level, so 1% significant at the 1% level, 5% at the 5% level, and 10% at the 10% level. Here we see that the observed percentages are much greater than these uniform percentages, in all cases statistically significantly different in a two-tailed test at any conventional level.

Given that authors are predisposed to report less significant results for sniff tests yet here we see more significant results than expected from a uniform distribution, one might be tempted to conclude that there is no evidence of publication bias—quite the reverse. However, we do not know that the underlying distribution of the test statistics would be uniform in the absence of publication bias. It could be that the typical study is badly balanced, badly specified, or shows strong pre-trends. Indeed, column (1) forces us to conclude that the outcome of the typical test is worse for the authors than would be expected from random statistical noise. How much worse cannot be determined. It could be that publication bias counteracts what would have even been even more significant results than shown in the column. We only know that publication bias is not strong enough to reverse the outcome that the typical sniff test is significant as much as six times more often that would be expected at random.

The kernel density plot in Panel (1) of Figure 3 spikes for the lowest p-values. The mountain of extra mass for p-values below 0.05 is taken from the $(0.05, 0.2)$ interval. Above 0.2, the density

Jann's (2005) add-on module for Stata, `kdens`. The default method in this module to account for bounds on supports is renormalization, which takes the standard kernel density estimate and divides it by the amount of local kernel mass lying inside the bounds of the support. See Jones (1993) for a detailed description. We use this default as well as the default Epanechnikov kernel.

is nearly collinear with the dotted horizontal line having height 1, the benchmark uniform density.

## 5.3. Results for Pure Subsample of RCT Balance Tests

A cleaner test of publication bias is provided by column (2) of Table 2, which focuses on the subsample of RCT balance tests that is pure in the sense of not involving the p-value improving measures of re-randomization, stratification, or matching. For this subsample, in the absence of publication bias, p-values should be uniformly distributed, with the percentage of p-values significant at each threshold level equaling the threshold. This is the case with the most stringent threshold, with exactly 1.0% significant at the 1% level. For the other two, less-stringent thresholds, the percentages are less than the thresholds, with 4.1% significant at the 5% level, and 5.8% significant at the 10% level. The last 5.8% result is significantly different from the 10% threshold at the 1% level.

An $F$ test indicates that the three proportions jointly are different than the uniform benchmarks at the 1% level. The estimate $\hat{\lambda} = -0.017$ implies that on average across the three thresholds, the empirical proportions fall short of their respective benchmarks by 1.7 percentage points in level terms. The estimate $\hat{\rho} = -0.368$ implies that on average acrsoss the three thresholds, the empirical proportions fall short of their respective benchmarks by 36.8% in proportional terms. Both estimates of common departures across thresholds are significant at the 1% level.

Panel (2) of Figure 3 shows a block of missing mass relative to the uniform benchmark for p-values below 0.1. The kernel density plot rises quickly to the uniform benchmark, which it tracks closely except for the highest p-values in the interval $(0.9, 1.0]$. It appears that the block of mass missing from $[0, 0.1)$ has been shifted to $(0.9, 1.0]$. Of the densities generated by structural model in Figure 2, the closest resemblance is to those in the middle panel, which posit only publication bias ($b > 0$), no study flaws ($m = 0$). The resemblance between the empirical and model densities are indeed quite close.

## 5.4. Results for Other Subsamples

The remaining columns in Table 2 explore the distribution of p-values in subsamples besides the pure sample of RCT balance tests. Column (3) shows the results from RCT balance tests when a balance improvement method (re-randomization, stratification, and/or matching) was possibly or

definitively employed. The proportion of significant p-values is higher in column (3) than (2) for all three thresholds, indicating that balance was worse in the improved than the pure sample. A possible explanation is that authors applied improvement methods in the most problematic cases. These cases may have been so problematic that they could not be fully corrected, remaining worse even after improvement than in the pure sample.[10]

The remaining columns (4)–(6) in Table 2 analyze tests other than RCT balance. Unlike pure RCT balance tests, for which the counterfactual distribution of p-values in the absence of publication bias is known to be uniform, for the hodgepodge of other tests in columns (4)–(6), the counterfactual distribution is unknown to us econometricians.

Considering column (4), the results for the pure subsample of non-RCT balance tests sharply contrast those in column (2) for the pure subsample of RCT balance tests. Rather than observing proportions of p-values below the critical thresholds of 1%, 5%, and 10% as we did in column (2), we see proportions far greater than the threshold values. Evidently, a substantial proportion of studies are failing sniff tests because of specification or identification problems, enough to swamp any publication biases that would tend to reduce the proportion of these significant p-values. The failure is systematic enough to call into question claims made in individual articles that failure is due to random bad luck. We will present additional evidence on the validity of such claims below in Table 4. The kernel density estimate for this subsample in panel (4) of Figure 3 looks similar to that for the full sample in panel (1), expected because this subsample constitutes the bulk of the full sample. Both closely resemble the densities generated by the structural model in the bottom panel of Figure 2, all having a spike for the lowest p-values which drops below the uniform benchmark for slightly higher values still below 0.1, then increasing again, resulting in a hump in the region above 0.1. The remarkable parallel between these highly nonlinear densities provides confidence in the validity of the structural model. Structural estimation will later allow us to distill the relative contributions of study flaws and publication bias from the shape of this density.

Columns (5) and (6) present results from subsamples where a technique to improve p-values was definitively employed in the paper. Specifically, this technique is matching since, as discussed, this is the only technique of relevance for the sample of non-RCT-balance tests. Column (5) shows

[10] Another possibility is that reviewers hold authors who have already applied balance-improving techniques to less stringent standards, allowing a few more significant balance tests to be attributed to random chance.

the huge proportion of significant p-values in matched samples before matching takes place. It is no surprise that the proportions differ from the uniform benchmark given that problematic p-values motivated authors to apply matching in the first place. Still, it is remarkable to see just how large the proportion of failed sniff tests can be in a problematic subsample like this.

Column (6) shows that, after matching, the proportion of significant p-values is greatly reduced, returning to values much closer to the thresholds. Matching appears to be an effective method to improve p-values in originally problematic studies. Panel (6) of Figure 3 confirms this impression, showing that p-values below 0.5 have a density considerably below the uniform line, much of this mass shifted to a spike of p-values above 0.9. The exception to this pattern is the 2.6% of p-values in the $[0, 0.01)$ interval, which column (6) of Table 2 shows is statistically significantly different from the 1% that would be expected if p-values were uniformly distributed. Evidently, the most severe flaws in some articles could not be addressed by matching alone.

## 5.5. Differences Across Categories

Table 3 provides additional evidence on publication bias by analyzing differences across various categories. We focus the analysis on two subsamples. The first set of columns studies differences across categories in the pure subsample of RCT balance tests. As before, these are good candidates for focused study because the counterfactual distribution of p-values in absence of publication bias is known to be uniform, and there is little reason to think that this uniformity would not persist in divisions of the sample according to, say, the location in which the test is reported (body versus appendix) or the tier of the journal publishing the containing article. The second set of columns studies the pure sample of tests other than balance in RCTs. While we do not know the counterfactual distribution of p-values for this subsample, we can still use it to test for publication bias under the assumption that the counterfactual distribution is independent of where the test is reported in the paper or the tier of the publishing journal. The subsample in the second set of columns still focuses on pure tests, excluding observations where matching was involved.[11] We

[11] We excluded matched samples for several reasons. First, all the articles eventually substitute post-match for pre-match data, so the pre-match p-values do not bear on data used for actual analysis in any article. Second, as shown in Table 2, the distribution of p-values for the pre-match subsample is wildly different from any other subsample. Although the pre-match subsample is relatively small, including it might add undue noise, obscuring any clear findings. Having excluded the pre-match subsample, it seemed natural to focus on a pure sample by excluding any matching entirely.

estimate the difference across categories in two ways, as a level effect $\Delta\hat{\lambda}$, applying IFWLS to the stacked regression (22), and as a proportional effect $\Delta\hat{\rho}$, applying IFWLS to the stacked regression (23). For space considerations, we only report the proportional differences $\Delta\hat{\rho}$ in Table 3; results for level differences $\Delta\hat{\lambda}$, which are qualitatively quite similar, are relegated to Appendix Table A2.

The table presents seven different specifications. Columns (1)–(5) examine category differences separately; column (6) combines them in a multiple regression; column (7) adds fixed publication-year and threshold effects to the multiple regression. The general pattern that holds for virtually all rows is that coefficients shrink in magnitude as more controls are added. The standard errors also decline, but more slowly than the coefficients, so the statistical signficance of the results decline as more controls are added. Indeed, the few significant results for pure RCT balance tests disappear in columns (6) and (7). The sample of 2,974 observations does not appear to provide sufficient power to detect the partial effect of category differences; none of the results are significant in columns (6) and (7) for pure RCT balance tests. On the other hand, the sample of other pure tests, with 17,175 observations, seems to provide enough power to detect some partial effects, so we will focus on that panel of the table, noting that the panel of results for pure RCT balance tests generally corroborate those for other pure tests regarding the sign and magnitude of results.

Consider the row corresponding to the appendix versus body difference. One might look for evidence of manipulation by authors in this comparison, hypothesizing that authors could try to obscure poor sniff tests by shunting them into the appendix, leading the appendix to contain a higher proportion of significant sniff tests than the body. The results provide no evidence of this form of manipulation. Especially in the columns (6) and (7) presenting the multiple regressions, the estimates of $\Delta\hat{\rho}$ are in fact negative, and in any event are small and statistically insignificant.

Consider the row corresponding to the omission versus inclusion of symbols highlighting statistical significance. One might look for evidence of manipulation by authors in this comparison as well, hypothesizing that authors may avoid using symbols when faced with many or badly failing sniff tests as the symbols would emphasize these failures, leading tables marked by symbols to have a lower proportion of significant p-values. The results provide no evidence of this form of manipulation either. The estimates of $\Delta\rho$ in this row are positive, the opposite of that predicted by the manipulation hypothesis, and in any event are small and statistically insignificant.

Some significant results emerge in the comparison across journal tiers. Other pure tests published in tier-2 and tier-3 journals are on average more significant than in tier 1. While the difference between tier 3 and tier 1 is not statistically significant, the difference between tier 2 and tier 1 is. In the specification (7) with the most controls, other pure tests in the average tier-2 journal are 19.7% more significant than in the average tier-1 journal. One interpretation is that the higher standards at better-tier journals lead to rejection of some studies with poorly performing sniff tests, which fall to the next tier or into a file drawer. Contrasting evidence is provided in the set of results for pure RCT balance tests. There we see that tier-1 journals publish 36.4% more significant sniff tests than tier-3 journals, although this result only shows up in column (3), disappearing in the columns (6) and (7) with more controls, and also does not show up in the comparison between tier-2 and tier-1 journals. At best we can say that there is modest evidence that top journals select on better sniff tests. It is worth emphasizing that this sort of selection does not necessarily distort statistical inference at prestigious journals as would selecting main results of interest on the basis of their $t$-statistics. However, it does support the possibility that rigorous standards at top journals can exacerbate publication bias.

The most substantial results show up in the last two rows. Large tables are those containing 12 or more sniff tests, the median number across tables. Columns (4) and (6) suggest the average large table has about twice the proportion of significant other pure tests than the average small table. Adding more controls in column (7) reduces the gap, but we still see that large tables exhibit 43.8% more significance than small. Several explanations are possible. It may be easier for authors to claim that significant sniff tests are statistical anomalies in large tables, or it may be more difficult for reviewers to focus on failed sniff tests in a complex paper with large tables. In any event, the correlation between table size and sniff-test performance justifies the need to apply inverse frequency weighting for unbiased estimation, as discussed in Section 4.2.

The last row of differences shows that tables in recent articles have less significant pure other tests than earlier articles. The dividing line for an article to be recent is being published from 2011 on, the midpoint of our sample period. This result is consistent with a decrease in study flaws for more recent articles. It is also consistent with an increase in publication bias over time.

## 5.6. Author Claims

Returning to the first results presented, those for the full sample in column (1) of Table 2, averaging across the three thresholds under study, we saw more significant p-values than would be expected under a uniform distribution by 4.2 percentage points in level terms and 45.1% in proportional terms. Panel (1) of Figure 3 shows a mountain of excess mass for p-values below 0.1. The probability that this departure from uniformity for the population of sniff tests is due to random bad luck is vanishingly small. In any individual case, however, authors may have the incentive to attribute unfavorable sniff tests to random bad luck, and it would be hard to dispute such individual claims. To investigate whether authors are overly willing to attribute unfavorable sniff tests to bad luck, we asked our research assistants to read the discussions of sniff tests in the articles and rate the authors' confidence in the test result according to a rubric involving four categories: "strong claim", "weak claim", "no claim" and "failed claim." We then analyzed the distribution of p-values for these categories. The proportions are reported in Table 4 and the differences in proportions across author claims of different strengths are reported in Table 5.

The "strong claim" category covers the case in which authors make an explicit or implied claim that unfavorable sniff-test results are due to random chance, not underlying imbalance or pre-treatment effects. If this is the case, we should observe no worse than a uniform distribution of p-values for this subsample and observe the proportion of significant p-values not to exceed threshold values. Looking at the pure sample of RCT balance tests analyzed in column (1) of Table 4, the strong claim seems to be justified. The proportion of p-values is significantly less than each of the three threshold values. The claims find some further justification in the pure sample of other tests analyzed in column (3). The proportions are lower than their corresponding 0.05 and 0.10 thresholds. The one place we see overclaiming is for the 0.01 threshold, with twice the proportion of p-values significant at the 1% level than would be expected at random, a difference significant at the 1% level. This suggests that, authors make overly strong claims for non-RCT-balance tests, especially so when dealing with the most stringent 1% threshold.

The "weak claim" category covers the case in which authors acknowledge the potential for imbalance or pre-treatment problems but go on to argue it does not completely undermine identification of their main results. Given that authors are acknowledging unfavorable sniff-test results, we would not expect the p-values to follow a uniform distribution. Indeed they do not. Column

(4) of Table 4 shows that the proportions greatly exceed threshold significance levels. For the "no claim" category, authors provide no discussion of the sniff test. Here again, we see from column (5) that the proportions of p-values exceed each of the three respective threshold values. For the last category, "failed claim," authors directly acknowledge that the sniff test failed and offer no defense of their identification strategy. The distribution of p-values that drive authors to make this admission is quite unfavorable indeed as shown in column (6).

Table 5 formally compares the differences across categories of author claims. The only difference that may run counter to expectations is that the "no claim" category has 72.2% fewer significant p-values on average across the thresholds than the "weak claim" category. The difference between the two categories of claim is large but too noisy to be statistically significant. The weak claim appears to be invoked as a defense against relatively poor sniff-test results, and can be taken by readers as a negative signal of test performance in the population of articles compared to when no claim is made—a feeble negative signal since the difference between "weak claim" and "no claim" is not statistically significant. The other notable findings from Table 5 are as expected. The performance of sniff tests in the "strong claim" category are significantly better, and "failed claim" category significantly worse, than for any of the other claim categories.

## 5.7. Threshold Effects

So far, we have uncovered evidence of publication bias but no evidence that it stems from author manipulation as opposed to some automatic processes generating the "file-drawer" problem. To investigate possible author manipulation further, Figure 4 presents an extremely fine-grained histogram of p-values from the full sample of observations that report exact values. The bin size is 0.001, implying that the figure displays 500 bins in total (were the horizontal axis not truncated for legibility, the figure would display 1,000 bins). If authors deliberately manipulated their data, methods, or results to reduce the significance of their sniff tests, we should observe a disproportionate share of p-values on or slightly above the 0.01, 0.05, and 0.10 thresholds.[12] The bars at almost every 0.01 tick mark are taller than the surrounding ones. This is due to benign rounding rather than devious manipulation as evidenced by the fact that the bars at these tick marks are as

[12] More precisely, the disproportionate share of p-values should show up exactly on the thresholds if manipulating authors exclude the threshold from the interval of significance and should show up slightly above if manipulating authors include the threshold in the interval of significance.

high in the 0.20 to 0.50 range—where there is no gain from manipulation because these thresholds have no reporting significance for such high p-values—as they are in the 0 to 0.10 range where thresholds have reporting consequences. There are no disproportionately high bars on or slightly above the thresholds of 0.01, 0.05, and 0.10. We treat the lack of disproportionately high bars around these thresholds as "precise zeros" in what amounts to a visual test for manipulation. We thus conclude from the figure that there was little or no effort to manipulate p-values from sniff tests to beat the thresholds for significance, at least in the subsample of observations included in the figure, for which the convention was to report exact p-values.

# 6. Structural Estimation

The structural model in Section 4.3 introduced parameters measuring the extent of publication bias ($b$) and study flaws ($m$). We can obtain structural estimates of these parameters by taking the model directly to the data. Equation (12) provides a density function that can be used in the construction of a log-likelihood function. For any subsample under consideration, some of the observations $i$ will have exact p-values $p_i$ reported; for others all we can glean is the interval $[\ell_i, u_i)$ containing $p_i$. The following log-likelihood function includes components for both sorts of observations:

$$\ln L = \sum_{p_i \text{ exact}} \ln f(p_i; m, b, \hat{\alpha}) + \sum_{p_i \in [\ell_i, u_i)} \ln \int_{\ell_i}^{u_i} f(p; m, b, \hat{\alpha})\, dp \tag{24}$$

$$\begin{aligned} = {} & N_{p_i<\hat{\alpha}} \ln(1-b) - N \ln(1 - \hat{\alpha}^{\frac{1}{1+m}} b) - \sum_{p_i \text{ exact}} \left[ \ln(1+m) + \frac{m}{1+m} \ln p_i \right] \\ & + \sum_{p_i \in [\ell_i, u_i)} \ln \Big\{ (1-b) \left[ \min(u_i, \hat{\alpha})^{\frac{1}{1+m}} - \min(\ell_i, \hat{\alpha})^{\frac{1}{1+m}} \right] \\ & \qquad + \max(u_i, \hat{\alpha})^{\frac{1}{1+m}} - \max(\ell_i, \hat{\alpha})^{\frac{1}{1+m}} \Big\}, \end{aligned} \tag{25}$$

where $\hat{\alpha}$ denotes the cutoff below which there is potential removal, $N$ denotes the number of observations in the subsample under consideration, and $N_{p_i<\hat{\alpha}}$ the number of those reporting an exact p-value that happens to be less than 0.1. Maximum likelihood estimates denoted with tildes, $\tilde{b}$ and $\tilde{m}$, can be obtained by maximizing $\ln L$. The inverse-frequency weights $w_i$ used in the reduced-form analysis will also be applied here in the maximum-likelihood procedure; see Section 5.1 for a discussion of the rationale. Based on visual inspection of the kernel densities in Figure 3, we set

$\hat{\alpha} = 0.1$ to avoid having to estimate this additional parameter.

Table 6 reports the structural estimates. For the pure sample of RCT balance tests, we obtain $\tilde{b} = 0.449$ and $\tilde{m} = 0.054$. The extremely low value of $\tilde{m}$, which is not statistically different from 0, implies that the pure sample of RCT balance tests exhibits almost no systematic failures of randomization; the distribution of p-values would be nearly uniform in the absence of publication bias. The value of $\tilde{b}$ is quite large, suggesting that publication bias removes 44.9% of p-values in the significant $[0, 0.1)$ interval.

These estimates can be linked to the reduced-form results from Table 2. Assuming for simplicity that all observations report exact p-values, it can be shown that

$$\lim_{m \to 0} \tilde{b} = \frac{\hat{\alpha} - \hat{\pi}_{\hat{\alpha}}}{\hat{\alpha}(1 - \hat{\pi}_{\hat{\alpha}})}, \tag{26}$$

where recall $\pi_{\hat{\alpha}}$ is the reduced-form estimate of the proportion of significant p-values below the cutoff $\hat{\alpha}$. Substituting $\hat{\alpha} = 0.1$ and the estimate $\pi_{\hat{\alpha}} = 0.058$ from column (2) of Table 2 yields

$$\frac{\hat{\alpha} - \hat{\pi}_{\hat{\alpha}}}{\hat{\alpha}(1 - \hat{\pi}_{\hat{\alpha}})} = \frac{0.1 - 0.058}{0.1(1 - 0.058)} = 0.446, \tag{27}$$

within a few hundredths of the structural estimate, $\tilde{b} = 0.449$. That the limit in (27) is so close to $\tilde{b} = 0.449$ justifies our claim that the estimate $\tilde{m} = 0.054$ is virtually 0, or in words that there are virtually no study flaws in the sample of pure RCT balance tests.

Intuitively, the structural model identifies study flaws from features of the functional form: the spikes in the density function around 0 and $\hat{\alpha} = 0.1$ shown in the bottom panel of Figure 2. Absent those spikes—which inspection of the kernel density in panel (2) of Figure 3 suggests are indeed absent—the estimate $\tilde{m}$ will be close to 0. With essentially no study flaws, the rate of removal is estimated from the amount of mass missing from below $\hat{\alpha} = 0.1$ regardless of precisely from where in the interval it is missing.

For the pure sample of tests besides balance in RCTs, we obtain $\tilde{b} = 0.913$ and $\tilde{m} = 3.906$. The estimates suggest pervasive study flaws in this subsample, generating huge mass of p-values in the significant $[0, 0.1)$ range that are then offset by a huge rate of publication bias. Published sniff tests truly represent the the "tip of the iceberg": less than 10% survive, with over 90% removed by publication bias.

Further confidence in the structural model can be obtained by comparing predicted to estimated densities. Consider the density predicted by the structural model presence of both publication bias and study flaws shown in the bottom panel of Figure 2. The density has a peculiar shape, spiking at $p = 0$, decreasing on $p \in (0, 0.1)$, jumping upward at $p = 0.1$, and decreasing again for yet higher p-values. The kernel density estimated for the sample of other pure tests in panel (4) of Figure 3 matches the predicted pattern. Contrast this shape with that of the kernel density estimated for pure RCT balance tests in panel (2) of Figure 3. Rather than having spikes at 0 and 0.1, this density has a block of mass missing from the interval below 0.1 but otherwise is fairly uniformly flat, resembling the predicted density with no study flaws, only publication bias, in the middle panel of Figure 2. It would be difficult to propose an alternative to our simple model delivering these non-monotonic patterns. It is not that the theory provides a rich portfolio of shapes that we sifted through to pick two that happen to match the empirical densities. Rather, the particular theoretical densities that match the corresponding empirical ones are predicated on conditions that fit the empirical settings. Balanced randomization is typically a much easier task and more likely to be successful than clean identification or correct specification. It is reasonable to suppose that pure RCT balance tests suffer from few study flaws compared to the specification, pre-trend, falsification tests, and non-RCT balance tests that constitute the category of other pure tests. In formal terms, it is reasonable to suppose $m \approx 0$ for the pure sample of RCT balance tests and $m > 0$ for the pure sample of other tests, the conditions behind the middle and bottom panels in Figure 2 matching the empirical densities for the respective samples.

# 7. Conclusion

In this paper, we shifted the traditional focus on the statistical significance of published papers' main results toward the p-values on auxiliary tests—including balance, over-identification, specification, falsification, and placebo tests—which we dubbed "sniff tests." We had two motives for studying sniff tests. First, since standards for passing are subjective, it is useful to have a picture of the distribution of sniff-test p-values in the broader economics literature to evaluate the performance of sniff tests in any individual paper. Second, since authors' have the reverse of the usual incentives regarding their p-values, preferring statistically insignificant values, sniff tests provide

a unique window on publication bias.

We estimated the distribution of p-values using a sample of nearly 30,000 sniff tests from 900 articles published in 60 economics journals. The distribution differed across subsamples of sniff tests. For the pure sample of RCT balance tests, there was little reason to expect randomization failures, and so the distribution of p-values under the null of no publication bias was expected to be uniform, allowing us to attribute departures from uniformity to publication bias. Our reduced-form estimate of the missing mass from the interval of significant p-values $[0, 0.1)$ is consistent with a rate of publication bias of around 45% (see equation (27)). That is, enough studies with significant pure RCT balance tests end up in the "file drawer" (either thrown out or published in a lower-tier journal outside of our sample) that there are 45% fewer significant such tests than expected. Our structural estimates were exactly in line with these descriptive, reduced-form results, finding $\tilde{m} = 0.054$, consistent with negligible study flaws (so negligible randomization failures in pure RCTs), and $\tilde{b} = 0.449$. Re-expressed in percentage terms, this latter estimate is within a percentage point of the rate of publication bias from the reduced-form estimates.

For the pure sample of other tests besides RCT balance, we saw a different pattern. Instead of missing mass in the $[0, 0.1)$ interval of significant p-values, we saw a spike of excess mass near 0 dropping down into a trough around 0.1, consistent with the presence of both a high rate of study flaws offset by publication bias removing mass from the significant interval. Our structural estimates, identified off non-monotonicities in the density of p-values, uncovered a huge rate of latent publication bias, as much as 91%, operating alongside a high rate of study flaws. If this quantitative result can be believed, it means that virtually all of this sort of sniff test are in articles that end up in the "file drawer" with only 9% being published in the tiers of journal in our sample.

Probing into the mechanisms behind the publication-bias process, reassuringly, we found no evidence of author manipulation. Sniff tests placed prominently in the body of the paper were no less significant than those potentially obscured in the appendix. Sniff tests with significance highlighted by symbols were no less significant than those lacking the symbols. A fine-grained histogram showed no extra mass of exact p-values right above thresholds that might be expected if authors fudged results or mined data to obtain desirable, insignificant results for their sniff tests.

Comparing p-values across journal tiers in our sample, we found modest evidence of less significant sniff tests in top than lower-tier journals, perhaps the result of top-tier journals maintaining

higher publication standards. While a laudable outcome for the profession, this selectivity suggests that the journals hierarchy can contribute to publication bias, with the best journals publishing results that are most likely to be selected on significance. There is no evidence that publication bias is diminishing in the latter half of our sample; if anything we see more missing mass among significant other pure tests more recently. An alternative explanation of the time trend is that studies are improving over time, suffering from fewer specification and other flaws.

It deserves emphasizing that "publication bias" means something different for the sniff tests studied here than the usual tests. "Publication bias" here refers to we mean a divergence between the distribution of p-values between conducted and published studies. Inferences drawn from sniff tests are not necessarily biased by selectivity. However, evidence of selectivity in sniff tests suggests the scope of publication bias where it would be more problematic for inference—for tests of the main results. Our results suggest that the process of publication bias leads to the removal of a chunk of results from publication, as high as 91% for some subsamples. If a similar chunk is removed from tests of central interest, there we may be getting a dangerously large proportion of statistically significant results that are an artifact of noise, and the danger may be particularly severe for the best journals.

Our analysis of how authors characterize their results suggests some care in interpreting author claims. Faced with the most significant sniff tests, with p-values in the $[0, 0.01)$ interval, authors tend to make overly strong claims that the results are due to statistical noise. The proportion of p-values in that interval is more than double what would be expected under a uniform distribution. We also saw that when authors make a "weak claim" about a sniff test, defending the validity of the study in despite significant p-values, if anything this is a worse signal about the performance of the sniff tests than when the author makes no claims about the results, suggesting that weak claims should be read with skepticism.

# References


*2015 Journal Citation Reports*. Thomson Reuters, 2015.

Altman, Douglas G. and J. Martin Bland. (1995) "Absence of Evidence is not Evidence of Absence," *British Medical Journal* 311: 485.

Andrews, Isaiah and Maximilian Kasy. (2017). "Identification of and Correction for Publication Bias," National Bureau of Economic Research working paper no. 23298.

Ashenfelter, Orley, Colm Harmon, and Hessel Oosterbeek. (1999) "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias," *Labour Economics* 6: 453–470.

Athey, Susan and Guido W. Imbens. (2017) "The Econometrics of Randomized Experiments," in Esther Duflo and Abhijit Banerjee, eds., *Handbook of Economic Field Experiments*, vol. 1, Elsevier, 73–140.

Bax, Leon and Karel G. Moons. (2011) "Beyond Bublication Bias," *Journal of Clinical Epidemiology* 64: 459–462.

Begg, Colin B. and Madhuchhanda Mazumdar. (1994) "Operating Characteristics of a Rank Correlation Test for Publication Bias," *Biometrics* 50: 1088–1101.

Begg, Colin B. and Jesse A. Berlin. (1988) "Publication Bias: A Problem in Interpreting Medical Data," *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 151: 419–463.

Brodeur, Abel, *et al*. (2016). "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics* 8: 1–32.

Bruhn, Miriam and David McKenzie. (2009) "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics* 1: 200–232.

Card, David and Alan B. Krueger. (1995) "Time-Series Minimum-Wage Studies: A Meta-Analysis," *American Economic Review Papers and Proceedings* 85: 238–243.

Christensen, Garret. S. and Edward Miguel. (2016) "Transparency, Reproducibility, and the Credibility of Economics Research," National Bureau of Economic Research working paper no. 22989.

DeLong, J. Bradford and Kevin Lang. (1992) "Are all Economic Hypotheses False?" *Journal of Political Economy* 100: 1257–1272.

Dickersin, Kay, Yuan-I. Min, and Curtis L. Meinert. (1992) "Factors Influencing Publication of Research Results: Follow-up of Applications Submitted to Two Institutional Review Boards," *Journal of the American Medical Association* 267: 374–378.

Dickersin, Kay, *et al.* (1987) "Publication Bias and Clinical Trials," *Controlled Clinical Trials* 8: 343–353.

Doucouliagos, Chris. (2005) "Publication Bias in the Economic Freedom and Economic Growth Literature," *Journal of Economic Surveys* 19: 367–387.

Driessen, Ellen, *et al*. (2015) "Does Publication Bias Inflate the Apparent Efficacy of Psychological Treatment for Major Depressive Disorder? A Systematic Review and Meta-Analysis of US National Institutes of Health-funded Trials," *PLoS One* 10: e0137864.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. (2008) "Using Randomization in Development Economics Research: A Toolkit," in T. Schultz and John Strauss, eds., *Handbook of Development Economics*, vol. 4, Elsevier, 3895–3962.

Duval, Sue and Richard Tweedie. (2000) "Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis," *Biometrics* 56: 455–463.

Easterbrook, Philippa J., *et al*. (1991) "Publication Bias in Clinical Research," *Lancet* 337: 867–872.

Egger, Matthias, *et al*. (1997) "Bias in Meta-Analysis Detected by a Simple, Graphical Test," *British Medical Journal* 315: 629–634.

Görg, Holger and Eric Strobl. (2001) "Multinational Companies and Productivity Spillovers: A Meta-Analysis," *Economic Journal* 111: 723–739.

Havranek, T. (2015). "Measuring Intertemporal Substitution: The Importance ofMethod Choices and Selective Reporting," *Journal of the European Economic Association* 13: 1180–1204.

Hedges, Larry V. (1992) "Modeling Publication Selection Effects in Meta-Analysis," *Statistical Science* 7: 246–255.

Hopewell, Sally, *et al*. (2009) "Publication Bias in Clinical Trials due to Statistical Significance or Direction of Trial Results," *Cochrane Database of Systematic Reviews*, issue 1, article no. MR000006.

Ioannidis, John and Chris Doucouliagos. (2013) "What's to Know about the Credibility of Empirical Economics?" *Journal of Economic Surveys* 27: 997–1004.

Ioannidis, John, T. D. Stanley, and Hristos Doucouliagos. (2017) "The Power of Bias in Economics Research," *Economic Journal* 127: F236–F265.

Iyengar, Satish and Joel B. Greenhouse. (1988) "Selection Models and the File Drawer Problem," *Statistical Science* 3: 109–117.

Jann, Ben. (2005) "KDENS: Stata module for univariate kernel density estimation," Statistical Software Component S456410, Department of Economics, Boston College. http://ideas.repec.org/-c/boc/bocode/s456410.html.

Jones, M. C. (1993) "Simple Boundary Correction for Kernel Density Estimation," *Statistics and Computing* 3: 135-146.

Krakovsky, Marina. (2004) "Register or Perish," *Scientific American* 291: 18–20.

McCrary, Justin, Garret Christensen, and Daniele Fanelli. (2016) "Conservative Tests under Satisficing Models of Publication Bias," *PloS One* 11: e0149590.

Moher, David, *et al*. (2010) "CONSORT 2010 Explanation and Elaboration: Updated Guidelines for Reporting Parallel Group Randomised Trials," *British Medical Journal* 340: c869.

Nelson, Jon P. (2014) "Estimating the Price Elasticity of Beer: Meta-Analysis of Data with Heterogeneity, Dependence, and Publication Bias," *Journal of Health Economics* 33: 180–187.

Peters, Jamie L., *et al*. (2006) "Comparison of Two Methods to Detect Publication Bias in Meta-Analysis," *Journal of the American Medical Association* 295: 676–680.

Rosenthal, Robert. (1979) "The 'File Drawer' Problem and Tolerance for Null Results," *Psychological Bulletin* 86: 638.

Rothstein, Hannah R., Alexander J. Sutton, and Michael Borenstein, eds. (2006) *Publication Bias in Meta-analysis: Prevention, Assessment and Adjustments*. New York: Wiley.

Schulz, Kenneth F., Douglas G. Altman, and David Moher. (2010) "CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials," *BMC Medicine* 8: 1.

Sellke, Thomas, M. J. Bayarri, and James O. Berger. (2001) "Calibration of p Values for Testing Precise Null Hypotheses," *American Statistician* 55: 62–71.

Stanley, T. D. (2005) "Beyond Publication Bias," *Journal of Economic Surveys* 19: 309–345.

Stanley, T. D. (2008) Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection," *Oxford Bulletin of Economics and Statistics* 70: 103–127.

Stanley, T. D. and Hristos Doucouliagos. (2014) "Meta-regression Approximations to Reduce Publication Selection Bias," *Research Synthesis Methods* 5: 60–78.

Stanley, T. D. and Stephen B. Jarrell. (1989) "Meta-Regression Analysis: A Quantitative Method of Literature Surveys," *Journal of Economic Surveys* 3: 161–170.

Stern, Jerome M. and R. John Simes. (1997) "Publication Bias: Evidence of Delayed Publication in a Cohort Study of Clinical Research Projects," *British Medical Journal* 315: 640–645.

Sterne, Jonathan A. C., David Gavaghan, and Matthias Egger. (2000) "Publication and Related Bias in Meta-Analysis: Power of Statistical Tests and Prevalence in the Literature," *Journal of Clinical Epidemiology* 53: 1119–1129.

Sterling, Theodore D. (1959) "Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa," *Journal of the American Statistical Association* 54: 30–34.

Wooldridge, Jeffrey M. (2010) *Econometric Analysis of Cross Section and Panel Data*, second edition. Cambridge, Massachusetts: MIT Press.

Table 1: Subsample Breakdowns

| | | | Refined breakdown | |
|---|---|---|---|---|
| | Number of observations | Percent of total sample | Number of observations | Percent of subsample |
| By methodology | | | | |
| • RCT balance tests | | | | |
| ○ Pure | | | 2,981 | 42% |
| ○ Possibly pure | | | 1,007 | 14% |
| ○ Improved | | | 3,162 | 44% |
| ○ Subtotal | 7,150 | 24% | 7,150 | 100% |
| • Other tests | | | | |
| ○ Pure | | | 17,210 | 76% |
| ○ Matched, pre-match sample | | | 1,886 | 8% |
| ○ Matched, post-match sample | | | 3,566 | 16% |
| ○ Subtotal | 22,662 | 76% | 22,662 | 100% |
| • Total | 29,812 | 100% | | |
| By precision of p-value gleaned | | | | |
| • Exact value | 12,157 | 41% | | |
| • Interval | 17,655 | 59% | | |
| • Total | 29,812 | 100% | | |
| By position in article | | | | |
| • Body | 21,570 | 72% | | |
| • Appendix | 8,242 | 28% | | |
| • Total | 29,812 | 100% | | |
| By symbols highlighting significance | | | | |
| • Symbols used | 19,410 | 65% | | |
| • Not used | 10,402 | 35% | | |
| • Total | 29,812 | 100% | | |
| By journal tier | | | | |
| • Tier 1 | 5,867 | 20% | | |
| • Tier 2 | 15,759 | 53% | | |
| • Tier 3 | 8,186 | 27% | | |
| • Total | 29,812 | 100% | | |
| By author claim about sniff test | | | | |
| • Strong claim | 21,260 | 71% | | |
| • Weak claim | 2,727 | 9% | | |
| • No claim | 1,959 | 7% | | |
| • Failed claim | 3,866 | 13% | | |
| • Total | 29,812 | 100% | | |

Notes: Breakdowns of final dataset consisting of 29,812 observations for which an exact value or interval could be gleaned for the p-value.

Table 2: Proportion of Significant P-values in Various Subsamples

| | | RCT balance tests | | Other tests | | |
|---|---|---|---|---|---|---|
| | Full sample | Pure | Possibly improved | Pure | Matched, pre-match | Matched, post-match |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Proportion of significant p-values, $\hat{\pi}_\alpha$ | | | | | | |
| • $\alpha = 0.01$ | $0.061^{***}$ | $0.010$ | $0.021^{*}$ | $0.053^{***}$ | $0.472^{***}$ | $0.026^{**}$ |
| | $(0.004)$ | $(0.004)$ | $(0.006)$ | $(0.005)$ | $(0.034)$ | $(0.007)$ |
| • $\alpha = 0.05$ | $0.097^{***}$ | $0.041$ | $0.056$ | $0.091^{***}$ | $0.562^{***}$ | $0.041$ |
| | $(0.005)$ | $(0.007)$ | $(0.008)$ | $(0.006)$ | $(0.033)$ | $(0.009)$ |
| • $\alpha = 0.10$ | $0.128^{***}$ | $0.058^{***}$ | $0.085$ | $0.124^{***}$ | $0.629^{***}$ | $0.060^{***}$ |
| | $(0.005)$ | $(0.008)$ | $(0.009)$ | $(0.006)$ | $(0.031)$ | $(0.011)$ |
| Combining results across thresholds | | | | | | |
| • Joint $F$-test | $68.6^{***}$ | $15.4^{***}$ | $8.29^{***}$ | $38.9^{***}$ | $96.8^{***}$ | $17.9^{***}$ |
| • Linear departure from uniform, $\hat{\lambda}$ | $0.042^{***}$ | $-0.017^{***}$ | $0.001$ | $0.036^{***}$ | $0.500^{***}$ | $-0.011$ |
| | $(0.004)$ | $(0.006)$ | $(0.008)$ | $(0.005)$ | $(0.031)$ | $(0.009)$ |
| • Proportional departure from uniform, $\hat{\rho}$ | $0.451^{***}$ | $-0.368^{***}$ | $-0.082$ | $0.393^{***}$ | $6.670^{***}$ | $-0.343^{***}$ |
| | $(0.060)$ | $(0.091)$ | $(0.106)$ | $(0.074)$ | $(0.402)$ | $(0.127)$ |
| Observation counts | | | | | | |
| • Sample size | $29,711$ | $2,974$ | $4,110$ | $17,175$ | $1,886$ | $3,566$ |
| • Stacked observations | $82,484$ | $7,895$ | $12,022$ | $47,007$ | $5,354$ | $10,206$ |
| • Clusters | $1,387$ | $103$ | $141$ | $988$ | $97$ | $157$ |

Notes: Results are the weighted proportions $\hat{\pi}_\alpha$ of observations $i$ that have p-values less than the threshold level $\alpha$ in the row heading. Weights equal $w_i$. Results can equivalently be obtained as coefficients from the stacked IFWLS regression (19). Each sniff-test statistic $i$ in the sample may contribute several observations to the stacked regression, depending on the number of thresholds under study $\alpha \in R^* = \{0.01, 0.05, 0.10\}$ that overlap with the containing table's reporting convention, $R_{t(i)}$. Column (3) combines possibly pure and definitively improved samples. Standard errors, reported in parentheses below results, are clustered at the table level. Clustering correctly adjusts standard errors for having multiple thresholds stacked per observation. Significantly different from $\alpha$ in a two-tailed test at the $^{*}$ten-percent level, $^{**}$five-percent level, $^{***}$one-percent level.

Table 3: Difference Across Categories in Proportion of Significant P-values

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| Differences for RCT balance tests, pure | | | | | | | |
| • Appendix – Body | 0.392 | | | | | 0.221 | 0.014 |
| | (0.273) | | | | | (0.277) | (0.132) |
| • No symbols – Symbols | | −0.075 | | | | 0.051 | 0.024 |
| | | (0.174) | | | | (0.172) | (0.086) |
| • Tier 2 – Tier 1 | | | 0.130 | | | 0.140 | 0.064 |
| | | | (0.243) | | | (0.246) | (0.144) |
| • Tier 3 – Tier 1 | | | −0.364* | | | −0.221 | −0.129 |
| | | | (0.198) | | | (0.224) | (0.102) |
| • Large table – Small table | | | | 0.342* | | 0.240 | 0.117 |
| | | | | (0.177) | | (0.190) | (0.103) |
| • Recent – Early | | | | | −0.464* | −0.358 | |
| | | | | | (0.277) | (0.265) | |
| Differences for other tests, pure | | | | | | | |
| • Appendix – Body | −0.133 | | | | | −0.207 | −0.098 |
| | (0.173) | | | | | (0.166) | (0.081) |
| • No symbols – Symbols | | 0.200 | | | | 0.057 | 0.049 |
| | | (0.192) | | | | (0.191) | (0.100) |
| • Tier 2 – Tier 1 | | | 0.418** | | | 0.365** | 0.197* |
| | | | (0.184) | | | (0.182) | (0.114) |
| • Tier 3 – Tier 1 | | | 0.058 | | | 0.116 | 0.060 |
| | | | (0.198) | | | (0.201) | (0.107) |
| • Large table – Small table | | | | 0.960*** | | 0.908*** | 0.438*** |
| | | | | (0.146) | | (0.144) | (0.126) |
| • Recent – Early | | | | | −0.607** | −0.501** | |
| | | | | | (0.239) | (0.238) | |
| Fixed effects | | | | | | | |
| • Publication year | No | No | No | No | No | No | Yes |
| • Threshold $\alpha$ | No | No | No | No | No | No | Yes |

Notes: Proportional difference $\Delta\hat{\rho}$ estimated from regression (23). See Appendix Table A2 for analogous results for level differences $\Delta\hat{\lambda}$. Regression stacks all test of the indicated type as well as all thresholds $\alpha \in R^*$. The sample of 2,974 pure RCT balance tests generate 7,895 stacked observations and 103 clusters. The sample of 17,175 pure other tests generate 47,007 stacked observations and 988 clusters. Standard errors, reported in parentheses below results, are clustered at the table level. Significantly different from 0 in a two-tailed test at the *ten-percent level, **five-percent level, ***one-percent level.

Table 4: Proportion of Significant P-values by Strength of Author Claim

| | RCT balance tests, pure | | Other tests, pure | | | |
|---|---|---|---|---|---|---|
| | Strong claim | Non-strong claim | Strong claim | Weak claim | No claim | Failed claim |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Proportion of significant p-values, $\hat{\pi}_\alpha$ | | | | | | |
| • $\alpha = 0.01$ | $0.005^{***}$ | 0.063 | $0.021^{***}$ | $0.116^{***}$ | $0.089^{**}$ | $0.294^{***}$ |
| | (0.002) | (0.032) | (0.003) | (0.021) | (0.033) | (0.030) |
| • $\alpha = 0.05$ | $0.029^{***}$ | $0.149^{*}$ | 0.044 | $0.208^{***}$ | $0.169^{***}$ | $0.390^{***}$ |
| | (0.005) | (0.050) | (0.004) | (0.025) | (0.034) | (0.026) |
| • $\alpha = 0.10$ | $0.045^{***}$ | $0.202^{**}$ | $0.071^{***}$ | $0.282^{***}$ | $0.217^{***}$ | $0.474^{***}$ |
| | (0.010) | (0.036) | (0.011) | (0.025) | (0.020) | (0.041) |
| Combining results across thresholds | | | | | | |
| • Joint $F$-test | $24.5^{***}$ | 2.31 | $41.8^{***}$ | $17.1^{***}$ | $4.6^{***}$ | $73.3^{***}$ |
| • Linear departure from uniform, $\hat{\lambda}$ | $-0.027^{***}$ | $0.084^{*}$ | $-0.008^{**}$ | $0.147^{***}$ | $0.105^{***}$ | $0.333^{***}$ |
| | (0.004) | (0.038) | (0.004) | (0.021) | (0.031) | (0.026) |
| • Proportional departure from uniform, $\hat{\rho}$ | $-0.521^{***}$ | $1.280^{**}$ | $-0.246^{***}$ | $2.192^{***}$ | $1.470^{***}$ | $4.614^{***}$ |
| | (0.069) | (0.538) | (0.056) | (0.308) | (0.415) | (0.335) |
| Observation counts | | | | | | |
| • Sample size | 2,622 | 352 | 10,641 | 1,890 | 1,732 | 2,912 |
| • Stacked observations | 6,901 | 994 | 29,189 | 5,336 | 4,771 | 7,711 |
| • Clusters | 94 | 10 | 813 | 70 | 44 | 109 |

Notes: Notes from Table 2 apply. For RCT balance tests, pure, we combine weak claims, no claims, and failed claims into single "non-strong claim" category since there are too few clusters to reliably estimate them separately.

Table 5: Proportional Difference in Significance of P-values by Strength of Author Claims

| | Row heading − column heading proportional difference $\Delta\hat{\rho}$ | | |
|---|---|---|---|
| | Strong claim | Weak claim | No claim |
| RCT balance tests, pure | | | |
| • Non-strong claim | 1.801*** | | |
| | (0.512) | | |
| Other tests, pure | | | |
| • Weak claim | 2.439*** | | |
| | (0.311) | | |
| • No claim | 1.717*** | −0.722 | |
| | (0.414) | (0.512) | |
| • Failed claim | 4.861*** | 2.422*** | 3.144*** |
| | (0.338) | (0.453) | (0.530) |

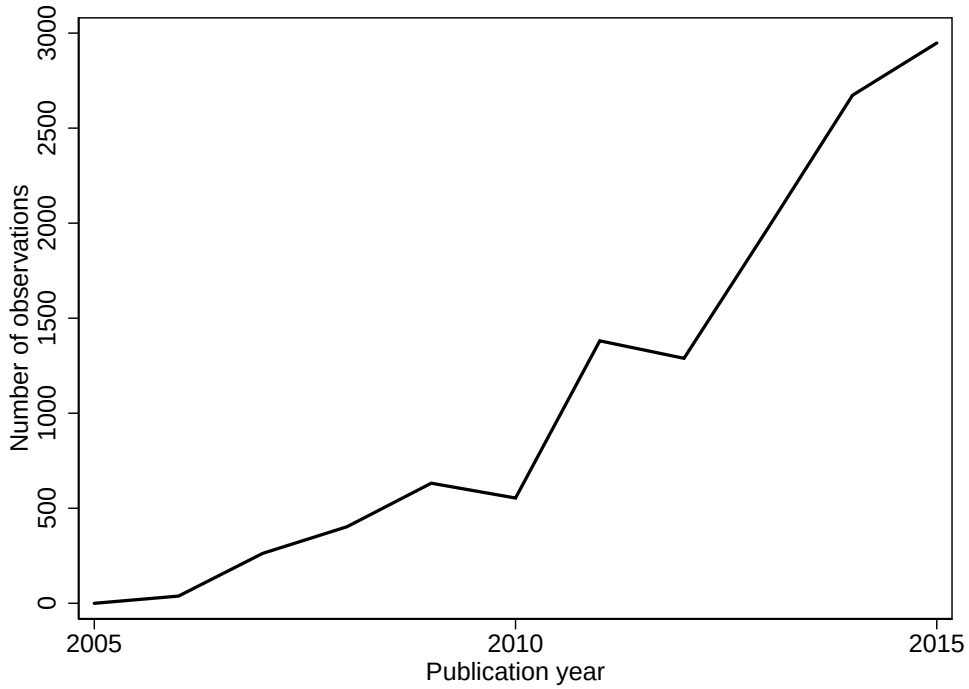
Notes: Proportional difference estimated from regression (23). Regression stacks all observations across author-claim categories as well as all thresholds $\alpha \in R^*$. Estimates can equivalently be computed by differencing $\hat{\rho}$ from the previous table. Entries above diagonal omitted for brevity; these have the same magnitude and opposite sign of the corresponding, reversed entries below the diagonal. See Appendix Table A3 for analogous results for level differences $\Delta\hat{\lambda}$. Standard errors, reported in parentheses below results, are clustered at the table level. Observation counts (including sample size, number of stacked observations, and number of clusters) can be computed from the previous table by summing the counts from the columns corresponding to the two categories being compared. Stars represent significant difference from 0 in a two-tailed test at the *ten-percent level, **five-percent level, ***one-percent level.

Table 6: Structural Estimates

| | RCT balance tests, pure | Other tests, pure |
|---|---|---|
| Structural parameters | | |
| • Publication bias, $b$ | $0.449^{***}$ $(0.132)$ | $0.913^{***}$ $(0.039)$ |
| • Study flaw, $m$ | $0.054$ $(0.107)$ | $3.906^{**}$ $(1.739)$ |
| Log-likelihood | $-403.2$ | $-4,024.3$ |
| Observations | 2,974 | 17,175 |
| Clusters | 103 | 988 |

Notes: Results from inverse-frequency-weighted maximization likelihood (IFWML). Likelihood function given in equation (25). Weights equal $w_i$. Significance cutoff for potential removal set to $\hat{\alpha} = 0.1$ rather than estimated. For the small number of observations whose p-values are smaller than machine double precision, producing an undefined natural logarithm, we rounded up to machine double precision; results nearly identical if these few observations are dropped. Standard errors clustered at the table $t$ level reported in parentheses. Stars indicate significant difference from 0 in a two-tailed test at the $^{*}$ten-percent level, $^{**}$five-percent level, $^{***}$one-percent level.

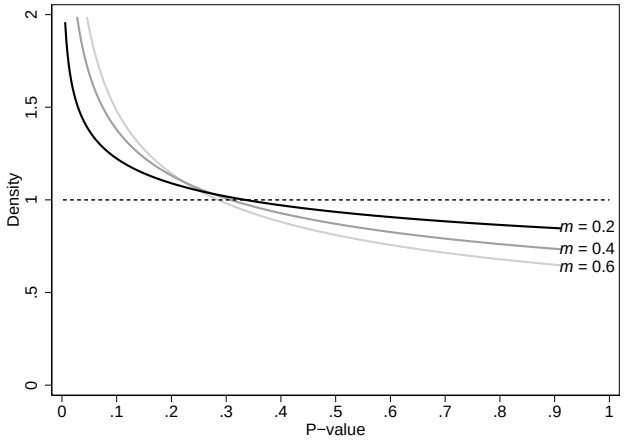
Figure 1: Trend in Sniff-Test Observations Over Time



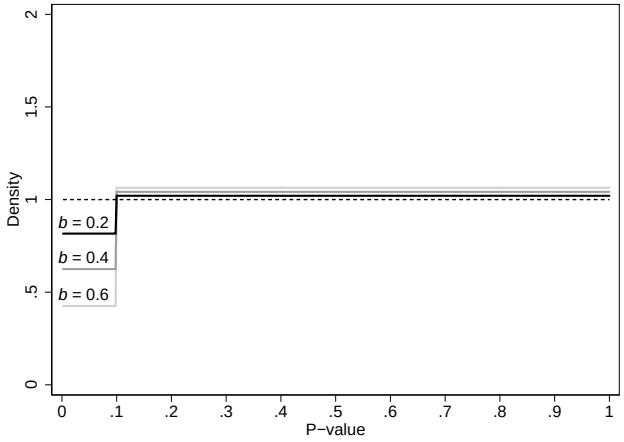

Note: Figure graphs the number of sniff-test observations in our dataset by the year that the containing article was published.

Figure 2: P-value Densities Generated by Model

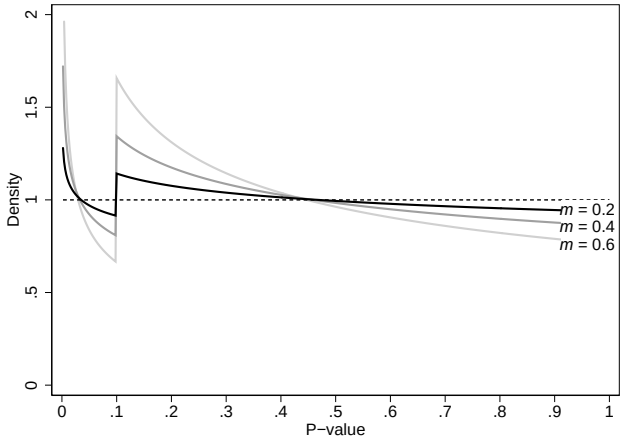
(1) Study flaws only, $b = 0$




(2) Publication bias only, $m = 0$

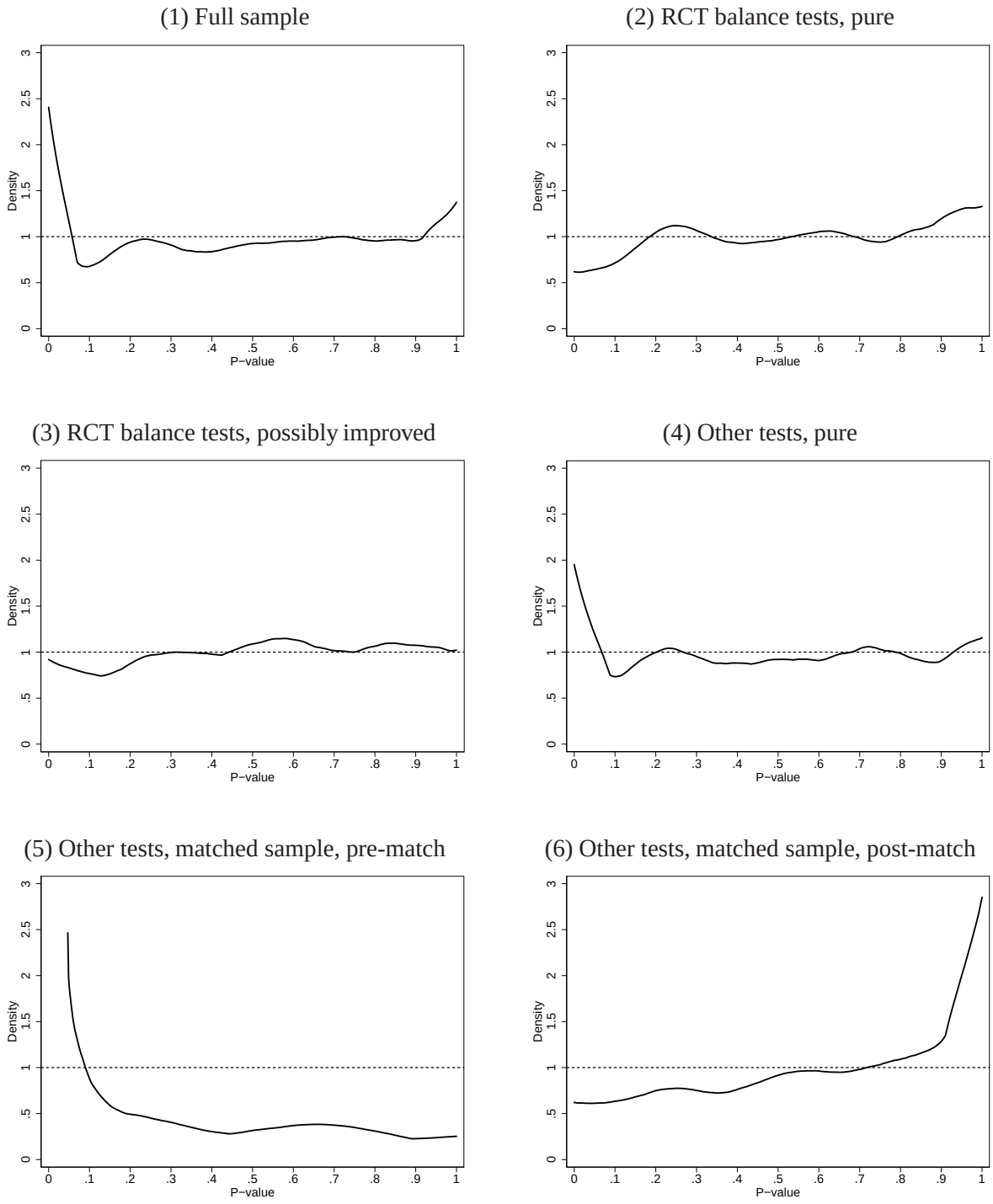



(3) Offsetting flaws and publication bias, $m = M(b, \hat{\alpha})$



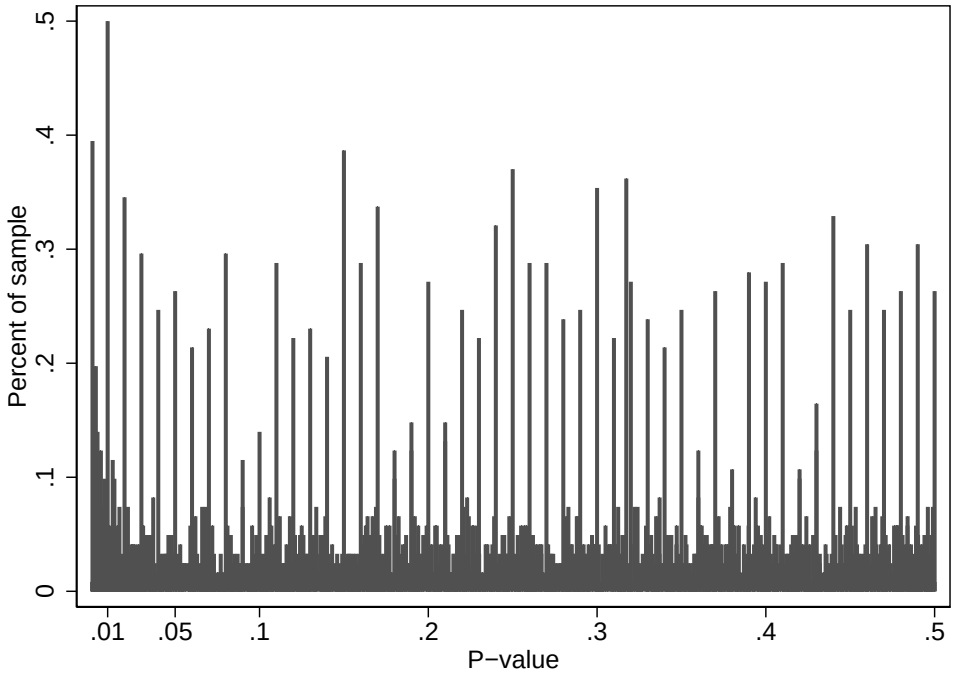

Note: Plots of density $f(p; m, b, \hat{\alpha})$ for $\hat{\alpha} = 0.1$ and for various values of $m$ and $b$. Dotted line is benchmark uniform distribution, corresponding to $m = b = 0$.

Figure 3: Kernel Density Estimates of Distribution of P-values




Notes: Each panel plots the kernel density estimate for the subsample of p-values analyzed in the corresponding column of Table 2. Solid curve is kernel density, estimated using Jann's (2005) `kdens` Stata module, accounting for lower and upper bounds on the support using the renormalization described in Jones (1993) and specifying an Epanechnikov kernel. To maintain consistent axes across panels, disproportionately high curve in panel (5) truncated at maximum label on the vertical axis. For comparison, dotted line is uniform $[0, 1]$ density.

Figure 4: Fine-Grained Histogram of Exact P-values




Note: Figure shows a fine-grained histogram, with bin size equal to 0.001, for the sample of observations for which exact p-values are reported. For legibility, only the distribution of p-values below 0.5 is shown and the disproportionately tall bar at p-value of 0 has been truncated at the maximum label on the vertical axis.

Table A1: Journals in Sample by Tier

| Tier 1 | | | Tier 2 | | | Tier 3 | | |
|---|---|---|---|---|---|---|---|---|
| Journal | Impact factor | Sample % | Journal | Impact factor | Sample % | Journal | Impact factor | Sample % |
| † *Quarterly J. Ec.* | 9.8 | 4.2 | *Ecological Ec.* | 3.9 | 1.5 | *J. Banking & Fin.* | 1.9 | 0.6 |
| *J. Fin. Ec.* | 5.9 | 1.9 | *Energy Ec.* | 3.4 | 0.1 | *J. Int. Money & Fin.* | 1.9 | 0.1 |
| † *Econometrica* | 5.8 | 2.5 | *J. Health Ec.* | 3.3 | 15.8 | *European J. Political Ec.* | 1.8 | 0.5 |
| † *J. Political Ec.* | 5.7 | 1.7 | *Ec & Human Biology* | 3.0 | 1.1 | *J. Corporate Fin.* | 1.8 | 0.4 |
| † *American Ec. Rev.* | 5.0 | 7.9 | *J. Envir. Ec. & Manag.* | 2.9 | 1.6 | *European Ec. Rev.* | 1.8 | 2.3 |
| † *Rev. Ec. Studies* | 4.7 | 1.5 | *J. Urban Ec.* | 2.9 | 3.9 | *China Ec. Rev.* | 1.8 | 1.1 |
| *J. Accounting & Ec.* | 4.7 | 0.1 | *Food Policy* | 2.8 | 0.0 | *J. Ec. Psychology* | 1.8 | 0.2 |
| | | | *J. Public Ec.* | 2.8 | 10.8 | *J. Comparative Ec.* | 1.7 | 1.3 |
| | | | *J. Development Ec.* | 2.8 | 10.3 | *J. Ec. Behavior & Org.* | 1.5 | 5.2 |
| | | | *J. Int. Ec.* | 2.7 | 1.1 | *Ec. Education Rev.* | 1.5 | 2.0 |
| | | | *World Development* | 2.7 | 5.8 | *J. Empirical Fin.* | 1.5 | 0.4 |
| | | | *J. Monetary Ec.* | 2.7 | 0.1 | *Regional Sci. & Urban Ec.* | 1.4 | 0.0 |
| | | | *J. Econometrics* | 2.3 | 0.4 | *Labour Ec.* | 1.4 | 8.3 |
| | | | *J. Fin. Stability* | 2.1 | 0.1 | *Int. Rev. Ec. & Fin.* | 1.4 | 0.5 |
| | | | *Resource & Energy Ec.* | 2.0 | 0.1 | *Int. J. Industrial Org.* | 1.4 | 0.6 |
| | | | | | | *Information Ec. & Policy* | 1.1 | 0.1 |
| | | | | | | *Explorations Ec. History* | 1.1 | 0.5 |
| | | | | | | *Pacific-Basin Fin. J.* | 1.0 | 0.0 |
| | | | | | | *J. Housing Ec.* | 1.0 | 0.0 |
| | | | | | | *Ec. Modelling* | 0.9 | 0.4 |
| | | | | | | *J. Japanese & Int. Ec.* | 0.8 | 0.1 |
| | | | | | | *Ec. Letters* | 0.7 | 0.9 |
| | | | | | | *Fin. Research Letters* | 0.7 | 0.0 |
| | | | | | | *Int. Rev. Law & Ec.* | 0.5 | 0.4 |
| | | | | | | *J. Applied Ec.* | 0.5 | 0.1 |

Note: Listing of journals constituting sample, divided into three tiers by five-year impact factor. Tier-1 journals have a five-year impact factor above 4, tier-2 from 2 to 4, and tier-1 below 2. Five-year impact factors from *Thomson Reuters Journal Citation Reports 2015*. All listed journals contribute observations to sample, though some entries in the % of sample column round to 0.0%. Table omits listing of the 12 journals in tier 3 that are too young to have a five-year impact factor by 2015: *Int. Rev. Ec. Education*, *Int. Rev. Fin. Analysis*, *J. Asian Ec.*, *J. Behavioral & Experimental Ec.*, *J. Choice Modelling*, *J. Fin. Intermediation*, *J. Socio-Ec.*, *North American J. Ec. & Fin.*, *Quarterly Rev. Ec. & Fin.*, *Research Policy*, *Research Social Stratification & Mobility*, and *Rev. Fin. Ec.*. None of these journals alone constitutes more than 0.5%; together, they constitute less than 1.5% of the sample. Journals accessed via JSTOR site designated by †; these journals also happen to be regarded as the top-five general interest journals in economics; journals without this designation are published by Elsevier and accessed via the ScienceDirect website.

Table A2: Level Difference in Significance of P-values Across Categories

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| Differences for RCT balance tests, pure | | | | | | | |
| • Appendix – Body | 0.015 | | | | | 0.006 | −0.002 |
| | (0.015) | | | | | (0.012) | (0.016) |
| • No symbols – Symbols | | −0.000 | | | | 0.007 | 0.006 |
| | | (0.011) | | | | (0.012) | (0.012) |
| • Tier 2 – Tier 1 | | | 0.007 | | | 0.007 | 0.005 |
| | | | (0.016) | | | (0.016) | (0.018) |
| • Tier 3 – Tier 1 | | | −0.019 | | | −0.013 | −0.015 |
| | | | (0.014) | | | (0.015) | (0.015) |
| • Large table – Small table | | | | 0.021* | | 0.018 | 0.017 |
| | | | | (0.011) | | (0.013) | (0.012) |
| • Recent – Early | | | | | −0.022 | −0.019 | |
| | | | | | (0.015) | (0.015) | |
| Differences for Other tests, pure | | | | | | | |
| • Appendix – Body | −0.010 | | | | | −0.015 | −0.012 |
| | (0.012) | | | | | (0.012) | (0.012) |
| • No symbols – Symbols | | 0.020 | | | | 0.012 | 0.011 |
| | | (0.014) | | | | (0.014) | (0.014) |
| • Tier 2 – Tier 1 | | | 0.032** | | | 0.029** | 0.029** |
| | | | (0.013) | | | (0.013) | (0.013) |
| • Tier 3 – Tier 1 | | | 0.004 | | | 0.008 | 0.008 |
| | | | (0.014) | | | (0.014) | (0.014) |
| • Large table – Small table | | | | 0.061*** | | 0.057*** | 0.057*** |
| | | | | (0.010) | | (0.010) | (0.010) |
| • Recent – Early | | | | | −0.038** | −0.030* | |
| | | | | | (0.016) | (0.016) | |
| Fixed effects | | | | | | | |
| • Publication year | No | No | No | No | No | No | Yes |
| • Threshold $\alpha$ | No | No | No | No | No | No | Yes |

Notes: Complementary results to those in Table 5 except estimates of difference in levels, $\Delta\hat{\lambda}$, rather than proportions, $\Delta\hat{\rho}$. Notes from that table apply with the following modification: level difference estimated using regression (22).

Table A3: Level Difference in Significance of P-values by Strength of Author Claim

| | Row heading − column heading level difference $\Delta\hat{\lambda}$ | | |
|---|---|---|---|
| | Strong claim | Weak claim | No claim |
| RCT balance tests, pure | | | |
| • Non-strong claim | 0.112*** | | |
| | (0.036) | | |
| Other tests, pure | | | |
| • Weak claim | 0.155*** | | |
| | (0.021) | | |
| • No claim | 0.133*** | −0.042 | |
| | (0.031) | (0.038) | |
| • Failed claim | 0.340*** | 0.185*** | 0.227*** |
| | (0.026) | (0.033) | (0.040) |

Notes: Complementary results to those in Table 5 except estimates of difference in levels, $\Delta\hat{\lambda}$, rather than proportions, $\Delta\hat{\rho}$. Notes from that table apply with the following modifications. Level difference estimated using regression (22). Estimates can equivalently be computed by differencing $\hat{\lambda}$ from Table 4.

# Appendix B: Supplementary Technical Details

This appendix fills in several technical details omitted from the text.

## Proof that P-Values Uniformly Distributed

Here we prove the claim in the text that $p_i$, the p-value defined in equation (4), has a uniform $[0,1]$ distribution under the null of no study flaw and under the further condition that there is no publication bias. We have

$$\begin{aligned} \Pr(p_i \leq \alpha | H_0) &= \Pr(\Gamma(p_i) \subseteq \Gamma(\alpha) | H_0) && \text{(B1)} \\ &= \Pr(\Gamma(A(Z_i)) \subseteq \Gamma(\alpha) | H_0) && \text{(B2)} \\ &= \Pr(\Gamma(A(Z_0)) \subseteq \Gamma(\alpha)) && \text{(B3)} \end{aligned}$$

Equation (B1) follows from the fact that $\Gamma(\cdot)$ is an increasing correspondence, as stated in (2). To see (B2), substituting $p_i$ for $\alpha$ in (1) yields $p_i = \Pr(Z_0 \in \Gamma(p_i))$. Combining this equation with (4), we have $p_i = A(Z_i)$, from which (B2) follows. Equation (B3) follows from imposing the null when drawing test statistic $Z_0$.

We proceed to bound (B3) above and below by $\alpha$. By the definition of $A(Z_i)$ in (3), $Z_0 \in \Gamma(A(Z_0))$. But then

$$\Gamma(A(Z_0)) \subseteq \Gamma(\alpha) \Longrightarrow Z_0 \in \Gamma(\alpha), \tag{B4}$$

which implies

$$\Pr(\Gamma(A(Z_0)) \subseteq \Gamma(\alpha)) \leq \Pr(Z_0 \in \Gamma(\alpha)) = \alpha, \tag{B5}$$

where the last equality holds by (1). To bound (B3) below by $\alpha$, note

$$\begin{aligned} \Gamma(A(Z_0)) \not\subseteq \Gamma(\alpha) &\Longrightarrow \Gamma(\alpha) \subset \Gamma(A(Z_0)) && \text{(B6)} \\ &\Longrightarrow Z_0 \notin \Gamma(\alpha). && \text{(B7)} \end{aligned}$$

Implication (B6) follows from the fact that $\Gamma(\cdot)$ is an increasing correspondence. as stated in (2), so that one of $\Gamma(A(Z_0))$ or $\Gamma(\alpha)$ must be included in the other. Implication (B7) follows from the fact that no smaller set than $\Gamma(A(Z_0))$ can include $Z_0$ by definition in (3). The chain of implications (B6)–(B7) implies

$$\begin{aligned} \Pr(\Gamma(A(Z_0)) \not\subseteq \Gamma(\alpha)) &\leq \Pr(Z_0 \notin \Gamma(\alpha)) && \text{(B8)} \\ &= 1 - \Pr(Z_0 \in \Gamma(\alpha)) && \text{(B9)} \\ &= 1 - \alpha, && \text{(B10)} \end{aligned}$$

where the last step follows from (1). But then

$$\begin{aligned} \Pr(\Gamma(A(Z_0)) \subseteq \Gamma(\alpha)) &= 1 - \Pr(\Gamma(A(Z_0)) \not\subseteq \Gamma(\alpha)) && \text{(B11)} \\ &\geq 1 - (1 - \alpha) && \text{(B12)} \\ &= \alpha, && \text{(B13)} \end{aligned}$$

where (B12) follows from (B8)–(B10). Equations (B5) and (B13) imply (B3) equals $\alpha$, implying $\Pr(p_i \leq \alpha | H_0) = \alpha$ by (B1)–(B3). This proves $p_i$ is uniformly distributed. □