EXAMINING SELECTION PRESSURES IN THE PUBLICATION PROCESS THROUGH THE LENS OF SNIFF TESTS

Christopher M. Snyder and Ran Zhuo*

Abstract—Economics papers increasingly report balance, pretrend, placebo, and other "sniff tests," rejection of which is bad news for authors, undermining the credibility of their main results. We derive nonparametric bounds on the latent proportion of significant sniff tests removed by the publication process (whether by *p*-hacking or relegation to the file drawer) and the proportion whose significance was due to true misspecification, not bad luck. Using a hand-collected sample of nearly 30,000 sniff tests, we estimate a removal rate of more than 30% for balance tests in randomized controlled trials and a misspecification rate of more than 40% for other tests.

I. Introduction

THE increasing demand for rigor in empirical economics has led to the increasing use of auxiliary tests (such as balance, parallel trends, overidentification, and placebo) to assess the credibility of a paper's main results. We dub these "sniff tests" because rejection is bad news for the author, and standards for passing are informal.

Sniff tests provide valuable information, allowing readers to discount main results that stem from misspecification. It is natural for journals not only to require sniff tests to be reported but also to use them as a screen, discarding misspecified studies that would pollute the literature and waste scarce journal space. The use of sniff tests as a screen is not completely benign, however. First, incentives to "publish or perish" may drive authors to omit sniff tests they have run indicating misspecification from their papers if not actively engaging in "*p*-hacking," manipulating data, methods, or reporting strategy so that reported probability values (*p*-values) appear better than actually estimated (Brodeur et al., 2020). These forms of author manipulation can distort results, impair statistical inference, and erode trust in research generally. Second, by chance, 5% of well-specified studies will have *p*-values that are significant at the 5% level, 10% significant at the 10% level, and so forth, making them appear to be misspecified when judged by those respective thresholds. If reviewers screen out these studies in the publication process (or authors self-screen anticipating reviewer treatment), valuable research would be lost, relegated to Rosenthal's (1979) metaphorical "file drawer."

In this paper we seek to measure the rate at which sniff tests are removed by the publication process (whether by *p*-hacking or relegation to the file drawer) and to determine whether this rate is justified by a commensurate rate of misspecification among removed studies. Estimating latent characteristics of removed tests, which are not directly observable, is a challenge, which we surmount by marshaling a large dataset of published sniff tests, hand collecting a sample of nearly 30,000 sniff tests from scores of economics journals. Under the null hypothesis of no misspecification, and absent removal by the publication process, sniff-test *p*-values should have a uniform distribution on [0, 1]. Comparing the aggregate distribution of *p*-values (called the *p*-curve) from our large sample of sniff tests to the uniform benchmark allows us to uncover the rate of removal of significant *p*-values and the extent of misspecification in the underlying set of papers.

Our analysis starts with a visual inspection of kernel-density estimates of *p*-curves in various subsamples. After an initial look at the full dataset, most of the analysis focuses on the "pure" sample of sniff tests for which the authors were unlikely to have taken measures (rerandomization, stratification, or matching) to lessen the *p*-values' significance. We further break the pure sample down into balance tests in randomized controlled trials (RCTs) and other tests. We expect that few RCTs suffer from flawed randomization procedures, ruling out the existence of a substantial proportion of misspecified studies for this subsample. Unless significant *p*-values have been removed by the publication process, the *p*-curve for balance tests in RCTs should follow the uniform distribution. For other tests, an unknown, perhaps substantial, proportion of studies may suffer from misspecification, skewing the *p*-curve toward 0 in the absence of removal.

For the pure sample of balance tests in RCTs, the kernel-density estimate of the *p*-curve looks relatively uniform except for a block of missing mass in the $[0, 0.15)$ interval. This shape is consistent with balance tests in RCTs not suffering from much misspecification but those with significant *p*-values being subject to a substantial rate of removal. For the pure sample of other tests, the *p*-curve is highly nonuniform, with a large spike of *p*-values near 0. This shape is consistent with substantial misspecification in the underlying population of studies. Proceeding from visual to formal regression analysis allows us to quantify the departure from the uniform benchmark and judge the statistical significance of any departure.

We proceed to use the regression estimates as an input into the computation of bounds on the rates of removal and

Received for publication October 19, 2021. Revision accepted for publication December 11, 2023. Editor: Xiaoxia Shi.

*Snyder: Dartmouth College and NBER; Zhuo: University of Michigan.

The authors are grateful to the Sloan Foundation for financial support and to Dartmouth College research programs for supplementary research-assistance funding. The paper greatly improved following the advice of editor Xiaoxia Shi and two referees. The authors also thank Barbara DeFelice, Ethan Lewis, Bruce Sacerdote, Douglas Staiger, Jonathan Zinman, and participants at the 9th Annual BITSS conference and Syms School of Business seminar for helpful suggestions and Kelsey Anspach, Ekshesh Bekele, Albert Chen, Barry Chen, Natalia Drozdoff, Isabel Hurley, Paul Jeon, Amrita Misha, Kirill Savolainen, and Olivia Zhao for dedicated research assistance.

A supplemental appendix is available online at https://doi.org/10.1162/rest_a_01410.

© 2024 The President and Fellows of Harvard College and the Massachusetts Institute of Technology
https://doi.org/10.1162/rest_a_01410

misspecification. One of our contributions in this paper is the derivation of such bounds under plausible nonparametric assumptions. Initially, we impose only the minimal assumption that misspecification tends to pile mass on low p-values whereas the publication process tends to do the opposite, disposing of low p-values or shifting them to higher values. Stronger assumptions allow us to tighten the bounds. Our tightest bounds suggest that around a third of the pure sample of RCT balance tests with p-values in $[0, 0.15)$ were removed by the publication process and more than 40% of the pure sample of other tests with p-values in $[0, 0.15)$ were indicative of misspecification rather than bad luck.

Of course, authors would like readers to believe that significant sniff tests are the result of bad luck rather than misspecification. Such claims can be made with impunity because they are difficult to reject on an individual basis. Our large sample allows us to evaluate whether such claims are justified in the aggregate. We find that authors defend approximately 80% of significant p-values as either passing or resulting from bad luck in the pure sample of RCT balance tests and approximately 50% in the pure sample of other tests. These rates of strong claiming are not inconsistent with estimates of the nonparametric bounds on misspecification but may be high enough to merit reader scrutiny.

The paper is organized as follows. Section II provides a literature review, and section III describes our dataset. Section IV provides some background econometric theory and derives our nonparametric bounds on removal and misspecification. Section V presents descriptive statistics and empirical estimates of the bounds. Section VI concludes. An extensive online appendix provides much of the technical detail behind the text including details on our sample construction, formal proofs of the propositions stated in the text, generalizations of those propositions, and formulas and empirical estimates for a suite of additional nonparametric bounds.

II. Relation to Literature

Our paper provides the first large-scale meta-analysis of sniff tests. The closest previous work is Bruhn and McKenzie (2009). The authors investigate the practice used by leading development economists to obtain and report balance in RCTs. The authors analyze a sample of balance tests from articles in development economics. Although their study examines thirteen 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 from sniff tests in the economics literature and to run formal statistical tests.

A series of more recent papers apply econometric theory to determine whether sniff tests can be appropriately used as a screen and to develop alternatives if not. Most of these papers focus on testing for violation of parallel trends in difference-in-difference studies, but the results often have more general implications. Kahn-Lang and Lang (2019) raised an early caution that an insignificant result from a parallel-trends test may not adequately justify the research design. Borusyak et al. (2022) propose new tests for pretrends in event studies. Andrews et al. (2020) propose a measure relating the performance of sniff tests to their informativeness on the robustness of main results. Freyaldenhoven et al. (2019) propose methods to estimate causal effects in event studies when pretrends are present in the outcomes. Instead of the null of no misspecification, Bilinski and Hatfield (2020) and Dette and Schumann (2024) suggest flipping the perspective and testing against the null that misspecification exceeds some threshold. Roth (2022) shows that screening studies based on violation of parallel trends can exacerbate publication bias in the main results of interest.

Much of the voluminous literature on publication bias focuses not on sniff tests but on main tests.[1] The broader literature on publication bias in medical and science journals is too vast to survey here. The literature on publication bias in economics, dating back at least to DeLong and Lang (1992), has been surveyed in Stanley (2005), Ioannidis and Doucouliagos (2013), and Christensen and Miguel (2018). Opportunities to identify the universe of unpublished and published studies is rare for meta-researchers in economics because the majority of economics studies are observational and preanalysis plans for RCTs have gained traction only recently. More commonly, only the selected set of published articles can be observed. To facilitate the detection of publication bias in this selected set, meta-analyses have focused on isolated cases in which many studies of the same pair of dependent and independent variables have been published, applying methods including the funnel plot, rank correlation tests, and parametric selection models. Examples include meta-analyses by Card and Krueger (1995) on the effect of minimum wage on employment, Ashenfelter et al. (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 Havránek (2015) on intertemporal substitution. Our methods allow us to pool observations across a range of topics, as do Brodeur et al. (2016) and Ioannidis et al. (2017).

Our approach shares some similarities with recent advances in this literature. Elliott et al. (2022) prove that in the absence of removal, the p-curve is nonincreasing; in the special case of t-tests, furthermore, the p-curve is completely monotonic and lies below specified bounds. Nonmonotonicities and excess mass above their specified bounds provide evidence of p-hacking. We also bound the p-curve, but our bounds are tailored to be dispositive for sniff tests, where p-hacking tends to remove significant p-values. Andrews and Kasy (2019) develop methods to nonparametrically identify the removal rate of insignificant main results using the joint distribution of coefficients and standard errors observed in replication studies and meta-analyses. Although they focus

[1] The medical literature is just beginning to recognize what Ioannidis (2024) dubs inverse publication bias, arising in clinical safety trials and other contexts in his taxonomy.

FIGURE 1.—TREND IN SNIFF-TEST OBSERVATIONS OVER TIME

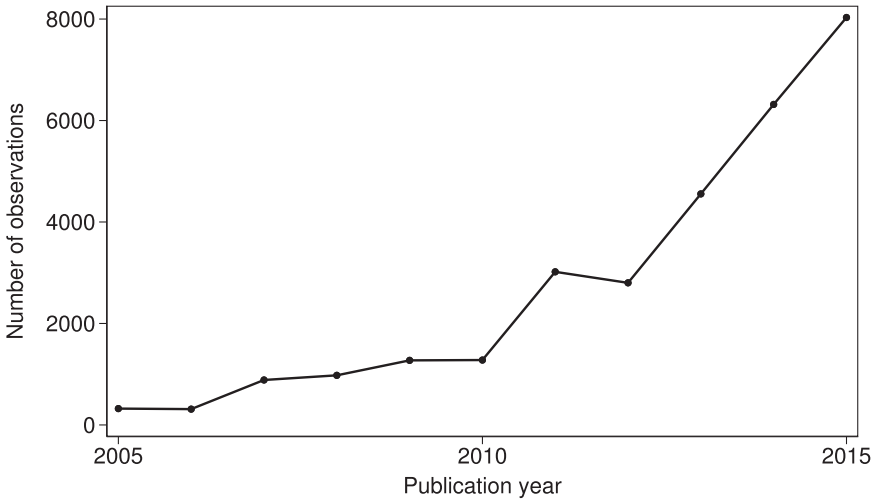


Figure graphs the number of sniff-test observations in our dataset by the year that the containing article was published.

on main results, their methods equally apply to identify the removal rate of significant sniff tests, the focus of our paper. Our alternative approach, tailored to the study of sniff tests, has some advantages in that context. We are able to bound the misspecification rate alongside the removal rate. Our approach can sometimes wring more usable observations from a sample, including observations for which authors do not report enough information to glean standard errors (e.g., reporting just significance thresholds) and observations reporting coefficients and standard errors to such low precision that computing a test statistic from them generates substantial rounding error. Differences aside, in the limited comparison conducted in the empirical section, the methods provide reassuringly similar results.

III. Data

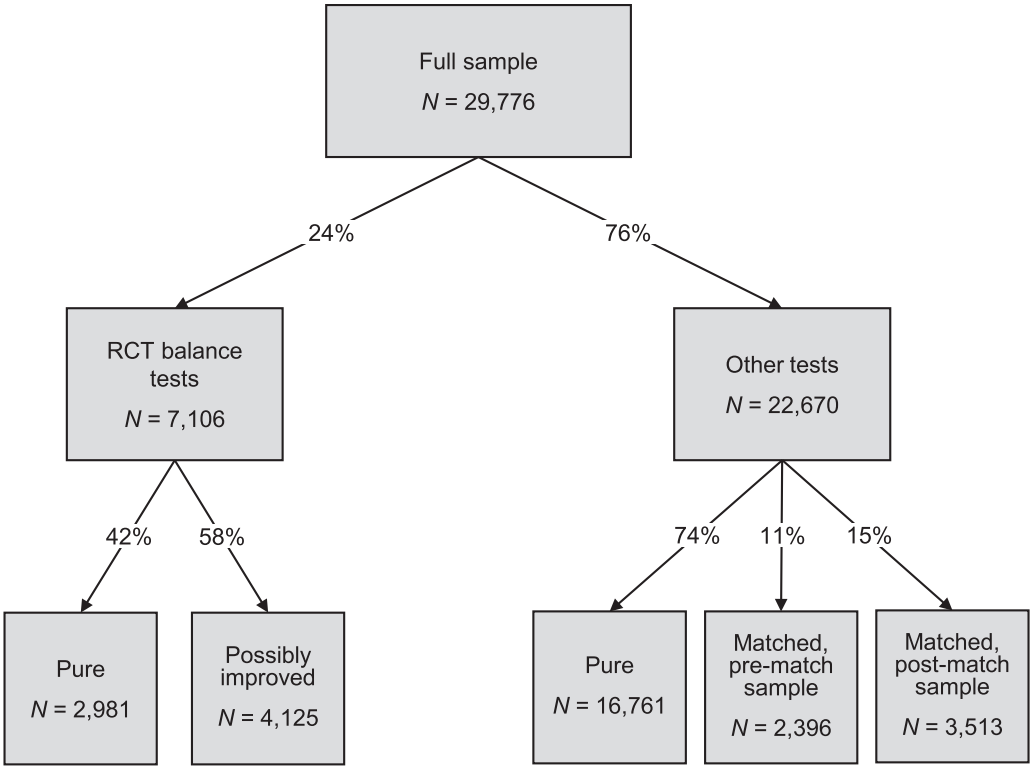
We collected data on sniff tests by having a team of research assistants systematically examine a large initial pool of journal articles in economics. We identified this pool from Elsevier's online database of journal articles, ScienceDirect. We collected all economics articles that were turned up by a search of related keywords such as "balance test," "falsification test," "placebo test," "randomization," "validation check," and so on (see the online appendix for a complete list). 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*, and *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 starting in 2005. The pool extends from 2005 to 2015.

The research assistants browsed each article in this initial pool, determining whether it contained a table reporting sniff tests. If so, the research assistants 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. All work was rechecked by supervising research assistants. In view of the difficulties of analyzing publication bias in the context of one-sided tests outlined by DeLong and Lang (1992), we collected only two-sided tests. We dropped 764 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 two further observations with *p*-values exceeding 1 leaves a final dataset of 29,776 sniff-test observations reported in 1,369 different tables from 891 articles published in 59 journals. A list of journals and the contribution each makes to the sample is provided in the online appendix. The *Journal of Health Economics* contributes the most observations (16% of the sample), followed by the *Journal of Public Economics* (11%), followed by the *Journal of Development Economics* (10%).

Figure 1 graphs the number of sniff-test observations in our sample by year of publication of the containing article, evincing a rapidly accelerating trend. Starting from a few observations in 2005, the number of sniff tests in our sample grows at a 38% compound annual rate.

Figure 2 breaks down our sample into subsamples by methodology. It is important to distinguish among methodologies because the distribution of *p*-values can differ systematically across them. A key breakdown is between tests of balance in randomized controlled trials (RCTs) and other tests. RCT balance tests, comprising 24% of our sample, play a special role in our analysis because what we presume to be the main source of misspecification with them—randomization failure—is presumably rare. In the absence of substantial misspecification, which could pile mass in the region of significant *p*-values, we have some hope of detecting missing mass in this region relative to the uniform [0, 1] distribution. As discussed in the theory

FIGURE 2.—SUBSAMPLE BREAKDOWN BY METHODOLOGY



section, under some conditions, missing mass relative to the uniform benchmark will allow us to bound the extent of study removal by the publication process.

The remaining 76% of our sample comprises other than RCT balance tests—including placebo tests, various falsification and specification tests, and balance tests in non-RCT settings. We do not have strong priors on the misspecification rate in these other tests. If many studies in our sample suffer from misspecification, the subsample of other tests may exhibit considerable excess mass at low p-values relative to the uniform benchmark.

Studies involving RCTs sometimes employ techniques to improve balance relative to what we will label the "pure" case of a single randomization. Athey and Imbens (2017) catalog the various balance-improvement techniques, including rerandomization (randomizing treatment/control selection multiple times until one achieves the desired balance in covariates between groups), stratification (randomizing treatment/control selection within covariate strata), and matching (finding pairs of observations with similar covariate values, making one a treatment and the other a control). Balance-improvement techniques can increase the power of tests of a study's main findings without biasing or distorting the size of those tests and indeed have been recommended to be used when possible as a principle of good study design (Morgan & Rubin, 2012).

The virtues of balance-improvement techniques aside, they do pose a challenge for our empirical approach because shifting some mass from low to high p-values creates a departure from the uniform distribution we use as a benchmark. To avoid misconstruing mass missing from the region of significant p-values due to balance improvement in published studies as indicating that studies have been removed by the publication process, we construct a "pure" subsample of RCT balance tests for which we deem balance improvement was unlikely to have been applied. We account for opaque or possibly missing mention of rerandomization by taking a conservative approach to constructing the pure subsample, including only items for which rerandomization was unlikely on a priori grounds. Among other examples, this includes cases in which a public lottery determines treatment status and cases in which treatment has been assigned by a third party. The opacity of the reporting of stratification and matching is less of a concern: owing to the deliberation involved in their application, we presume that they are reported whenever used. Of the subsample of RCT balance tests, 42% are classified as pure. For the remaining 58%, either (a) the authors did not mention rerandomization in the article but we cannot definitively rule out rerandomization because authors had access to the baseline data to rerandomize before treatment assignment or (b) the authors definitively employed a balance-improvement technique. We conservatively classify these observations as "possibly improved."

We similarly break the subsample of tests other than RCT balance down by whether balance-improvement techniques have been employed. Rerandomization and stratification are irrelevant for these observations because authors

do not have requisite control over the experimental procedure outside of an RCT. The remaining feasible technique is matching, which authors can employ by restricting analysis to a matched subsample of their full sample. In 74% of the subsample of other tests, no matching, and thus no p-value improvement method, was employed. We deem this subsample as the "pure" sample of other tests. The rest involve some matching. For 11% of other tests, matching is employed, but the reported sniff tests are those conducted prior to matching. For 15% of other tests, matching is likewise employed, but the reported sniff tests are those conducted after matching.

Most of our analysis will focus on the pure subsamples of both RCT balance tests and other tests. We will also briefly examine the subsamples in which balance was possibly improved to see how strongly p-curves can depart from the uniform benchmark when unbridled selection forces operate.

We motivated the breakdown of our sample into subsamples by the systematic differences in p-curves across methodologies. Another motivation for analyzing these subsamples separately is that the consequences of using the sniff test as a publication screen can differ across methodologies, benign in some cases and biasing inference in others. For studies involving RCTs, if publication process removes studies failing their balance tests, this typically does not distort the size of tests of main findings for published RCT studies. For non-RCT studies that use sniff tests to assess the validity of their causal-identification strategies, removing studies that fail the pretests can bias tests of main results and distort inference (Andrews, 2018; Roth, 2022). For both RCT balance and other tests, removal by the publication process can potentially reduce social welfare by relegating useful research to the proverbial file drawer, not publicly circulated. The social benefit from keeping misspecified studies out of circulation is likely to be greater with other tests than RCT balance tests, presuming that the main source of misspecification—failed randomization—is rare in RCTs.

For the sniff tests in our sample, we recorded the p-value information provided by the article, which in some cases reported exact p-values, and in other cases an interval for the p-value (often a reported significance level as indicated by asterisks alongside the reported test statistic). For articles providing no direct information about p-values, we tried to glean p-values from ancillary information provided in the sniff-test table (perhaps a reported test statistic, perhaps other reported results from which we could derive a test statistic). Our sample has exact p-values for 41% of the observations and an interval for 59% of them.

IV. Theory

In this section we begin by providing theoretical background on the distribution of p-values from sniff tests. We explain why the uniform distribution is a useful benchmark for sniff-test p-values and how the distribution is altered by the presence of misspecification and removal by the publication process. Subsequent subsections construct nonparametric bounds on removal and misspecification, the estimation of which in the results section is a key contribution of the paper.

A. *Asymptotic Distribution of Sniff-Test p-Values*

This subsection provides background on the distribution of p-values from sniff tests, drawing on textbook material from Lehmann and Romano (2005). Let parameter vector θ index the misspecification (of whatever form relevant to the application) that a representative sniff test is designed to detect. The sniff test is structured as a test of the null hypothesis $H_0 : \theta \in \Theta_0$ of no misspecification against the alternative $H_1 : \theta \in \Theta_1$ of misspecification, where $\Theta_0 \cap \Theta_1 = \emptyset$. The author computes the sniff-test statistic Z_θ (the distribution of which depends on the presence and extent of misspecification θ) and determines whether Z_θ lies in rejection region $R(\alpha)$, where α is a prespecified significance level. Rejection regions are sometimes based on the exact distribution of Z_θ, sometimes on the asymptotic distribution, and are typically structured so that they are nested:

$$R(\alpha') \subset R(\alpha'') \text{ for } \alpha' < \alpha''. \tag{1}$$

The p-value associated with test statistic Z_θ is defined as the lowest significance level for which H_0 is rejected:

$$p \equiv \inf\{\alpha \mid Z_\theta \in R(\alpha)\}. \tag{2}$$

Let $\mathcal{F}(p, \theta)$ denote the cumulative distribution function (cdf) associated with the p-value.

Our benchmark result is that in the absence of distortions due to the publication process, the asymptotic distribution of the p-value from the sniff test is uniform on [0, 1] under the hypothesis H_0 of no misspecification; formally,

$$\mathcal{F}(p, \theta) = p \text{ for all } p \in [0, 1] \text{ and } \theta \in \Theta_0. \tag{3}$$

For a textbook proof and technical conditions required for equation (3) to hold, see, for example, Lehmann and Romano (2005, lemma 3.3.1). Notice that the uniform benchmark in equation (3) is not conditional on the number of observations n in the study employing the sniff test, holding whether the study involves few or many observations.

The presence of misspecification alters the distribution of p-values in a predictable way, piling more mass on lower p-values than the uniform benchmark. More formally, under weak conditions, the asymptotic distribution of p-values in the presence of of misspecification (but absent publication removal) is first-order stochastic dominated by the uniform distribution:

$$\mathcal{F}(p, \theta) \geq p \text{ for all } p \in [0, 1] \text{ and } \theta \in \Theta_1. \tag{4}$$

Equation (4) holds under the weak condition that the sniff test is unbiased, meaning that it is more likely to reject the

null if the alternative is true than if the null is true, or equivalently, that the sniff test's power exceeds its significance.[2]

The analysis so far has focused on representative sniff test. It is straightforward to see that the results (3) and (4) extend from an isolated sniff test to a collection of them, say, all the p-values reported in a table of sniff tests, whether or not the sniff tests in the collection are correlated. The results can be further extended to a sample of sniff tests from unrelated studies, such as in the dataset we will use for our empirical analysis. Let $F(p)$ denote the cdf averaged over the population of sniff tests, $F_0(p)$ the cdf averaged over the subpopulation for which the null of no misspecification holds, and $F_1(p)$ the cdf averaged over the subpopulation for which the alternative of some misspecification holds. Formally,

$$F(p) \equiv \int_{\Theta_0 \cup \Theta_1} \mathscr{F}(p, \theta) dG(\theta), \tag{5}$$

$$F_0(p) \equiv \frac{\int_{\Theta_0} \mathscr{F}(p, \theta) dG(\theta)}{\int_{\Theta_0} dG(\theta)}, \tag{6}$$

$$F_1(p) \equiv \frac{\int_{\Theta_1} \mathscr{F}(p, \theta) dG(\theta)}{\int_{\Theta_1} dG(\theta)}, \tag{7}$$

where $G(\theta)$ denotes the cdf of θ for the population of sniff tests. As the result (3) holds for all cdfs over which $F_0(p)$ is integrated, $F_0(p)$ inherits the result, implying $F_0(p) = p$. The result (4) holds for all cdfs over which $F_1(p)$ is integrated, implying $F_1(p) \geq p$. Together, the results (3) and (4) imply $F(p) \geq p$.

B. Nonparametric Bounds on Removal

The publication process may remove sniff tests with lower p-values in one of several ways. After seeing the significant result, the author may omit the sniff test from the table or omit the whole table. The author may decide to shelve the paper, or journals may reject it. Any of these mechanisms would *delete* the sniff test from the published record, reducing the number of published sniff tests. Other forms of removal may *shift* the distribution of sniff-test p-values in the published record up without reducing the total number. For example, an author may p-hack the sniff test to raise its p-value so that it appears less significant.

To formally model the ways in which the publication process can remove p-values, suppose the unit interval can be partitioned into two subintervals, $[0, r)$ and $[r, 1]$, where r is the threshold level of significance below which the publication process may exert pressure to remove p-values but above which it does not. We will refer to the lower interval $[0, r)$ as the removal region and the upper interval $[r, 1]$ as the no-removal region. Let $d_r \geq 0$ denote the mass of p-values deleted from the removal region and $s_r \geq 0$ the mass shifted from the removal to the no-removal region. We assume that mass is not deleted from the no-removal region nor is mass shifted back from the no-removal to the removal region. Assume that deletion and shifting from the removal region are exhaustive and mutually exclusive ways the publication process can remove sniff tests.

We next introduce a series of variables related the distribution of p-values in the pre- and postremoval populations of studies. Let m denote the mass of p-values in the preremoval population, π_r the proportion in the removal region of preremoval population, and π_n the proportion in the no-removal region of the preremoval population. Analogs of the variables just defined written with breves (a mnemonic for change, in this case resulting from removal) refer to their counterparts for the postremoval population. Thus, $\breve{m}$ denotes the mass of sniff tests in the postremoval population, $\breve{\pi}_r$ the proportion in the removal region in that population, and $\breve{\pi}_n$ the proportion in the no-removal region in that population. Normalize $m = 1$, implying that π_r and π_n represent both proportions and masses. Because the removal and no-removal regions partition each population, we have $\pi_r + \pi_n = \breve{\pi}_r + \breve{\pi}_n = 1$.

The variables introduced thus far in the subsection are population-level variables, equivalently, the theoretical probability limits to which consistent finite-sample estimators converge. Our dataset is a sample of the postremoval population, allowing us to obtain estimates, $\hat{\pi}_r$ and $\hat{\pi}_n$, of the respective population proportions, $\breve{\pi}_r$ and $\breve{\pi}_n$, with some error.

With that notation in hand, we can proceed to define the removal and misspecification rates we will proceed to estimate. Let ρ_r denote the removal rate of p-values from $[0, r)$, given by

$$\rho_r = \frac{d_r + s_r}{\pi_r}. \tag{8}$$

In principle, one could define an analogous removal rate ρ_n from $[r, 1]$, but because there is no removal in this region by assumption, we trivially have $\rho_n = 0$. Our analysis will thus focus on characterizing ρ_r. Ideally, we would be able to derive point estimates, but because ρ_r depends on unobserved properties of the preremoval population, absent functional-form assumptions we will have to be content with nonparametric bounds. Let $\underline{\rho}_r$ and $\bar{\rho}_r$ denote, respectively, lower and upper bounds on ρ_r such that no admissible configuration of removals $\{d_r, s_r\}$ exists for which ρ_r lies below $\underline{\rho}_r$ or above $\bar{\rho}_r$.

To be an admissible configuration of removals, $\{d_r, s_r\}$ must first satisfy the nonnegativity conditions $d_r \geq 0$ and $s_r \geq 0$. Further, they must satisfy the following identities to ensure that all p-values pre- and postremoval are accounted for, in every region and in the population as a whole. The identity

$$\breve{m} = m - d_r = 1 - d_r \tag{9}$$

[2] To prove equation (4), let $\beta(\alpha, \theta)$ denote the power of a sniff test with significance level α against alternative $\theta \in \Theta_1$, that is, $\beta(\alpha, \theta) = \Pr[Z_\theta \in R(\alpha)]$. Then $\mathscr{F}(\alpha, \theta) = \Pr_\theta(p \leq \alpha) \geq \Pr[Z_\theta \in R(\alpha)] = \beta(\alpha, \theta) \geq \alpha$. The first step follows from the definition of a cdf, the second step from the fact that $Z_\theta \in R(\alpha)$ implies $p \leq \alpha$ by equation (2), and the third step from definition of power. The final step holds for unbiased sniff tests. Exchanging the variable p for α yields equation (4).

ensures that mass in the whole population is accounted for. The identity

$$\breve{m}\breve{\pi}_r = \pi_r - d_r - s_r \tag{10}$$

ensures that mass in the removal region is accounted for.

We begin construction of bounds on ρ_r for the case in which the null hypothesis of no misspecification holds for all observations in a population. Absent misspecification, publication removal is the only force driving the p-value distribution away from the uniform benchmark, allowing us to use deviations of the postremoval population from the uniform benchmark to bound removal. Consider a hypothetical example in which all removal was due to shifting, not deletion. With no deletion, the preremoval mass of p-values is preserved postremoval, implying $\breve{m} = m = 1$ by equation (9). The proportion of p-values in the removal region (also the mass because the unit mass is preserved) differs by the amount of the shift: $s_r = \pi_r - \breve{\pi}_r$ by equation (10). Substituting the preceding equality along with $d_r = 0$ into equation (8) yields $\rho_r = (\pi_r - \breve{\pi}_r)/\pi_r$. Under the null of no misspecification, as shown in the previous subsection, $\pi_r = F_0(r) = r$, which upon substituting into the preceding equation yields

$$\rho_r = \frac{r - \breve{\pi}_r}{r}. \tag{11}$$

Continue to suppose that the null hypothesis of no misspecification holds for all observations, but now suppose that all removal was due to deletion, not shifting. Substituting $s_r = 0$ into equation (10), further substituting $\breve{m} = 1 - d_r$ from equation (9), and solving for d_r yields $d_r = (\pi_r - \breve{\pi}_r)/(1 - \breve{\pi}_r)$. This is precisely the deletion from the removal region needed so that the proportion of mass in that region is the same as the proportion observed in the postremoval population. Substituting this value of d_r along with $s_r = 0$ into equation (8) yields $\rho_r = (\pi_r - \breve{\pi}_r)/[\pi_r(1 - \breve{\pi}_r)]$. Under the null of no misspecification, $\pi_r = r$, and the preceding equation becomes

$$\rho_r = \frac{r - \breve{\pi}_r}{r(1 - \breve{\pi}_r)}. \tag{12}$$

It is easy to see that equation (12) exceeds equation (11). Attributing removal to shifting generates a lower removal rate than attributing it to deletion. Shifting has a dual effect on the departure of the postremoval distribution of p-values from the uniform benchmark, not only removing mass from $[0, r)$ but also piling the removed mass in $[r, 1]$. Deletion only has the first effect. Rationalizing a given departure of the postremoval population from the uniform benchmark therefore requires less shifting than deletion. Extending this logic, when removal is due to a combination of both shifting and deletion, the associated removal rate ρ_r will lie between equations (11) and (12). Because we have no other forms of removal besides shifting and deletion, we have established that, under the null of no misspecification, equation (11) provides a lower bound $\underline{\rho}_r$ on the removal rate and equation (12) provides an upper bound $\bar{\rho}_r$ on the removal rate.

The analysis has so far maintained the null of no misspecification. In the presence of possible misspecification, equation (11) still provides a valid lower bound. To see this, by equation (4), misspecification can only inflate the mass of p-values in $[0, r)$ relative to the uniform benchmark, requiring more removal to arrive at the postremoval mass. In the presence of misspecification, the upper bound in equation (12) is no longer valid, however. Examples can be constructed with increasingly extreme misspecification in which π_r is increasingly large, requiring an increasing removal rate to arrive at the observed postremoval distribution. As this construction suggests, the upper bound on the removal rate in the presence of possible misspecification is the trivial one, $\bar{\rho}_r = 1$.

C. *Nonparametric Bounds on Misspecification*

The preremoval proportion of p-values in the removal region, π_r, can be written as a weighted sum of p-values in that region from well-specified studies and p-values from misspecified studies:

$$\pi_r = F(r) = (1 - \mu)F_0(r) + \mu F_1(r), \tag{13}$$

where μ denotes the overall misspecification rate among studies in the preremoval population. We will focus on bounding not μ but μ_r, the misspecification rate restricted to the removal region. The reason that the publication process removes studies in $[0, r)$ is presumably to avoid the misspecification concentrated there. Quantifying μ_r will help clarify the tradeoffs involved in removal. The mass of misspecified studies in $[0, r)$ is given by the last term in equation (13). Dividing that term by the mass in $[0, r)$ expresses misspecification in that region as a rate:

$$\mu_r = \frac{\mu F_1(r)}{\pi_r}. \tag{14}$$

To derive a bound on μ_r, we will first distill some insights from equation (13). Noting that $F_0(r) = r$ as shown in section IVA and that $F_1(r) \leq 1$ because F_1 is an average of cdfs, substituting these expressions into equation (13) and rearranging yields

$$\mu \geq \frac{\pi_r - r}{1 - r}. \tag{15}$$

Intuitively, the extra mass in the removal region relative to the uniform benchmark, suitably normalized, bounds the overall misspecification rate μ. To translate the bound on μ into a bound on μ_r, note that $\mu F_1(r) = \pi_r - (1 - \mu)r \geq (\pi_r - r)/(1 - r)$, where the equality follows from rearranging equation (13) and the inequality from equation (15). Substituting this inequality into the numerator of equation (14) yields

$$\mu_r \geq \frac{\pi_r - r}{(1 - r)\pi_r}. \tag{16}$$

Suppose that there is no removal, that is, $\rho_r = 0$. Then equations (9) and (10) imply $\pi_r = \breve{\pi}_r$, which upon substituting into equation (16) yields

$$\mu_r \geq \frac{\breve{\pi}_r - r}{(1-r)\breve{\pi}_r}. \tag{17}$$

The right-hand side of equation (17) provides a lower bound $\underline{\mu}_r$ on the misspecification rate in the removal region absent removal.

Regardless of what we assume about the removal rate, we will not be able to derive a tighter upper bound of the misspecification rate than the trivial $\bar{\mu}_r = 1$. Intuitively, even if the preremoval *p*-value distribution appeared uniform, we cannot rule out that virtually all observations were misspecified (i.e., $\mu = 1$) but so slightly that the *p*-value distribution hardly departs from the uniform [so $F_1(r) \approx r$]. By equations (13) and (14), that would imply $\mu_r = 1$ as well.

The analysis so far has supposed that there is no removal. The possibility of removal only enlarges the scope for misspecification. Whatever the departure from the uniform distribution in the postremoval population, there may have been a greater departure preremoval, requiring even more misspecification to explain. The lower bound $\underline{\mu}_r$ in equation (17) continues to be valid with more potential misspecification. The upper bound was at the corner before, $\bar{\mu}_r = 1$, and cannot expand beyond that with more potential misspecification.

D. *Refining the Bounds*

The previous two subsections provided bounds on removal and misspecification under special circumstances. For example, we derived bounds on removal assuming either that there is no misspecification or that the extent of misspecification is unknown. Similarly, we derived bounds on misspecification assuming either that there was no removal or the extent of removal is unknown.

In this subsection we refine the bounds, allowing us, the researchers, to have partial information on misspecification in bounding removal or partial information on removal when bounding misspecification. The bounds are provided in the following series of propositions, proved in the online appendix.

Proposition 1. *Suppose we, the researchers, know* $\mu_r \leq \bar{\mu}_r$. *For all admissible removal configurations* $\{d_r, s_r\}$, *the removal rate* ρ_r *does not fall below the lower bound*

$$\underline{\rho}_r = \frac{r - \breve{\pi}_r}{r} \tag{18}$$

or above the upper bound

$$\bar{\rho}_r = \frac{r - \breve{\pi}_r + \breve{\pi}_r \bar{\mu}_r}{r(1-\breve{\pi}_r)}. \tag{19}$$

Proposition 2. *Suppose we, the researchers, know* $\rho_r \geq \underline{\rho}_r$. *For all admissible removal configurations* $\{d_r, s_r\}$, *the misspecification rate* μ_r *does not fall below the lower bound*

$$\underline{\mu}_r = \frac{\breve{\pi}_r - r + (1-\breve{\pi}_r)r\underline{\rho}_r}{(1-r)\breve{\pi}_r} \tag{20}$$

or above the (trivial) upper bound $\bar{\mu}_r = 1$.

Beyond providing new results, propositions 1 and 2 nest all the results derived in the preceding two subsections. Thus, a rigorous foundation can be provided for the sketches of derivations in the preceding two subsections. For example, the upper bound equation (12) on removal in the absence of misspecification can be recovered from proposition 1 by substituting $\bar{\mu}_r = 0$ into equation (19). Other upper and lower bounds (whether with or without information on complementary rates) can be recovered in a similar way.

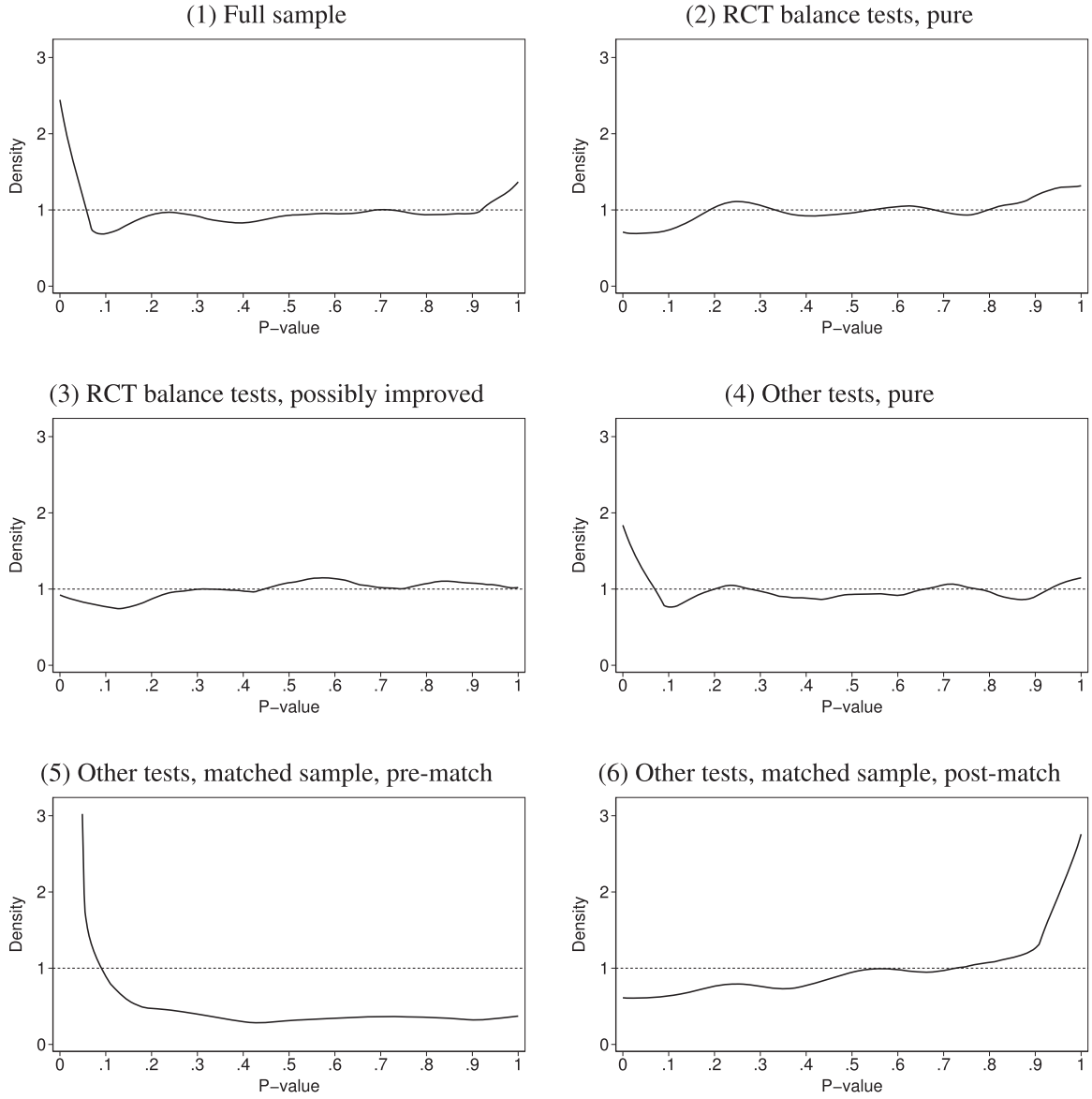
A bound is sharp if it can be attained for some admissible removal configuration, which has the key implication that it cannot be improved upon. The bounds in propositions 1 and 2 may or may not be sharp. We prove in the online appendix that the bounds on removal indeed are sharp in the absence of misspecification. We have proved that upper bound on removal equation (19) is not sharp by constructing an alternative bound that can improve upon it in rare circumstances. Given that the circumstances are not only rare but are not empirically relevant for our sample and that the improvement is so slight, we relegate discussion of the alternative bound to the online appendix. Some of the bounds can be slightly improved if, instead of aggregating the removal region into a single interval, it is divided into subintervals and the proportion of *p*-values in each subinterval separately estimated (as we do in the first three rows of table 1). The improvement is too slight to merit including a discussion in the main text of the rather complex recursive formulas involved (reflecting a procedure we call backward ironing), so we relegate the propositions and proofs in the case of multiple removal regions to the online appendix.

For consistency, we focused on the the removal region of the preremoval sample in bounding rates of removal and misspecification. Although these are arguably the most useful rates to bound, alternatives might be of interest as well. For example, one might be interested in bounding rates in the entire unit interval, not just the removal region, or one might be interested in bounding rates in the postremoval rather than the preremoval population. Owing to space constraints, we provide a comprehensive analysis of these alternatives in the online appendix.

V. Empirical Results

We next turn to the empirical results, presented in a series of subsections roughly in order of increasing rigor. Starting with a visual inspection of kernel density plots, we move to regressions used to conduct formal statistical tests of the

FIGURE 3.—KERNEL DENSITY ESTIMATES OF *p*-CURVES



Solid curve is the *p*-curve, estimated using Jann's (2005) KDENS Stata module with the default Epanechnikov kernel. Procedure accounts for lower and upper bounds on the support using the renormalization procedure prescribed by Jones (1993) taking the standard kernel-density estimate and dividing it by the amount of local mass lying inside the bounds of the support. To maintain consistent scaling across panels, disproportionately high curve in panel 5 is truncated at the maximum of vertical axis. Dotted line is the uniform [0, 1] density benchmark.

departures of the distributions from the uniform benchmark. A further subsection uses the regression estimates as inputs into estimation of the nonparametric bounds on removal and misspecification. The final subsection provides additional context for the bounds by bringing in additional data on the strength of author claims about how well their sniff tests performed.

A. Kernel-Density Estimates of p-Curves

Figure 3 plots kernel-density estimates of the *p*-curves from various samples of sniff tests. Standard methods for estimating kernel densities require data points rather than intervals, so the sample we use for the kernel densities includes only the 41% of observations for which we can glean exact *p*-values. To account for the fact that different tables contribute different number of sniff tests, we weight each *p*-value by the inverse of the number of observations in the containing table so each table contributes equally to the aggregate distribution. (Further details on the inverse-frequency-weighting method are provided in the next subsection.)

The *p*-curve in panel 1 is estimated using the full sample of exact *p*-values. Its peculiar nonlinear shape is a clear departure from uniformity. The mass spikes for the lowest *p*-values, suggesting a sizable number of published sniff

tests pick up true misspecification. The [0.05, 0.15) interval is missing mass relative to the uniform benchmark (the dotted line having height 1). Above 0.2, the density is close to the uniform benchmark.

Panel 2 plots the *p*-curve for the subsample of pure RCT balance tests. As discussed in previous sections, absent removal from the publication process, the distribution should be close to uniform because study flaws arising from failures to appropriately randomize treatment and control groups should be small. We observe a block of missing mass relative to the uniform benchmark for *p*-values below 0.15. Above that threshold the *p*-curve rises quickly to the uniform benchmark, which it tracks closely until it reaches the highest *p*-values in (0.85, 1]. The *p*-curve rises above the uniform benchmark in (0.85, 1], appearing as if the block of mass missing from [0, 0.15) was shifted there.

Panel 3 plots the *p*-curve for RCT balance tests when a balance improvement method (rerandomization, stratification, and/or matching) was definitively employed or likely employed. The proportion of significant *p*-values is higher than that in panel 2 and closely tracks the uniform benchmark, indicating an absence of detectable misspecification in this subsample and little additional removal of significant sniff tests.

Panel 4 plots the *p*-curve for the pure subsample of other tests. The similarity of this plot to that for the full sample in panel 1 is natural given that the subsample in 4 constitutes the majority of the full sample.

Panel 5 plots the *p*-curve for other tests conducted on the prematched sample by studies using matching. Authors usually show these *p*-values to motivate the use of matching to improve balance and these tests are usually accompanied by *p*-values of the same tests on the postmatched sample. Because the postmatched sample is used for estimation, the publication process likely exerts little pressure to remove significant sniff tests for prematched sample. Panel 5, therefore, provides a picture of unvarnished misspecification. We see a mountain of mass of *p*-values in [0, 0.1), after which the *p*-curve dips well below the uniform benchmark.

Panel 6 plots the *p*-curve for other tests, matched sample postmatch. We see quite the opposite of the prematch sample. *P*-values below 0.5 have a density considerably below the uniform benchmark, much of this mass shifted to a spike of *p*-values above 0.9, suggesting matching as an overall effective method to improve balance.

B. Regression Estimates of Interval Proportions

In this section, we quantify the qualitative observations made from the kernel-density plots using regression analysis. Another advantage of the regression analysis is that it can exploit a larger sample containing both exact *p*-values and some observations for which only intervals are reported.

A key choice in the specification of the subsequent analysis is how large to make the removal region. Our choice of [0, 0.15) is motivated by the estimates obtained in this subsection of proportions of *p*-values in different subintervals of length 0.05. To explain which data are included in the estimation of these proportions requires some additional notation. Let p_i denote a *p*-value observation, indexed by $i \in I$, where I denotes the population of *p*-values. Index tables by $t \in T$, where T denotes the population of tables. We sometimes emphasize the "containing" relation with functional notation, letting $t(i)$ denote the table containing p_i. We seek to measure the share of *p*-values falling into a given subinterval $j = [a_{j-1}, a_j) \subset [0, 1]$. Equivalently, letting $\mathbf{1}_{p_i \in j}$ be the indicator function for the subscripted event that p_i is contained in j, we seek to estimate the expected value of $\mathbf{1}_{p_i \in j}$.

One challenge in estimating this expected value is that tables differ in the precision at which they report *p*-values, some specifying an exact value for p_i, others reporting whether p_i achieves a single significance level, and others partitioning [0, 1] into multiple significance levels. The reporting convention used in some tables may not line up with j in a way that would allow us to deduce $\mathbf{1}_{p_i \in j}$ from the information provided by the table. Let T_j denote the subset of tables whose reporting convention does allow us to deduce $\mathbf{1}_{p_i \in j}$ for any realization of p_i.[3] For example, a table reporting significance at the 5% level is in $T_{[0,0.05)}$ but not in $T_{[0,0.01)}$ or in $T_{[0,0.10)}$. We include only tables in T_j in our analysis of interval j.[4] Let I_j denote the sniff tests in the subset of tables T_j.

The proportion of *p*-values that fall into interval j for a typical table is captured by the following conditional expectation:

$$\breve{\pi}_j = E_{t \in T_j}[E(\mathbf{1}_{p_i \in j} \mid i \in t)]. \tag{21}$$

The conditional expectation in equation (21) corresponds to the sampling frame that first randomly samples a table and then randomly samples a sniff test within that table, in effect allowing each table to contribute equally to our estimates. A natural alternative for the sampling frame would be to pool sniff tests across tables and sample directly from that pool. The drawback of this alternative is that tables with more sniff tests would tend to be overrepresented, biasing the estimates if the number of sniff tests presented in an article is correlated with their significance (see the online appendix for a proof).

[3] Formally, T_j is the set of tables whose reporting convention induces a partition of [0, 1] that is at least as fine as $\{j, [0, 1] \setminus j\}$, the coarsest partition containing j.

[4] Obviously, we cannot use observations outside of T_j for which $\mathbf{1}_{p_i \in j}$ cannot be determined. We could include observations outside of T_j for which $\mathbf{1}_{p_i \in j}$ can be determined, but doing so would impart a selection bias. For example, let $j = [0, 0.10)$. For a table reporting significance at the 5% level, we can determine $\mathbf{1}_{p_i \in j}$ for all significant *p*-values reported in the table but cannot determine $\mathbf{1}_{p_i \in j}$ for all insignificant *p*-values.

A consistent estimator $\hat{\pi}_j$ of the population proportion $\breve{\pi}_j$ can be obtained by replacing the expectations in equation (21) with sample averages:

$$\hat{\pi}_j = \frac{1}{|T_j|} \sum_{t \in T_j} \left(\frac{1}{|t|} \sum_{i \in t} \mathbf{1}_{p_i \in j} \right), \tag{22}$$

where vertical bars denote the cardinality of sets, so $|T_j|$ denotes the number of tables whose reporting convention line up with j and $|t|$ the number of sniff tests in table t. We can obtain a convenient, equivalent expression for $\hat{\pi}_j$ by introducing inverse frequency weights

$$w_i = \frac{|I_j|}{|T_j||t(i)|}. \tag{23}$$

These weights are inversely proportional to the number of sniff tests $|t(i)|$ in the included table, scaled by the constant $|I_j|/|T_j|$ required for the weights to sum to $|I_j|$, the number of relevant observations. Substituting equation (23) into equation (22) yields

$$\hat{\pi}_j = \frac{1}{|I_j|} \sum_{i \in I_j} w_i \mathbf{1}_{p_i \in j}, \tag{24}$$

the inverse-frequency-weighted average of $\mathbf{1}_{p_i \in j}$ for the relevant subsample of sniff tests.

Regression-based methods can be used to recover $\hat{\pi}_j$. In particular, in the inverse-frequency-weighted least squares (IFWLS) regression of $\mathbf{1}_{p_i \in j}$ on a constant using the subsample $i \in I_j$, that is,

$$\mathbf{1}_{p_i \in j} = \breve{\pi}_j \cdot 1 + u_{ij} \qquad i \in I_j, \tag{25}$$

the estimate $\hat{\pi}_j$ of the coefficient on the constant term is numerically identical to $\hat{\pi}_j$ in equation (22). Regression-based methods have the advantage of facilitating the computation of standard errors. Because tables form the sampling frame in our analysis, one natural clustering scheme is the table level. To account for correlation among sniff tests possibly constructed from the same dataset underlying an article, we adopt the more conservative approach of clustering at the article level for all reported standard errors.

Table 1 presents the regression estimates for a selection of the samples covered by panels in figure 3. The first column presents aggregate results for the full sample. We see that 9.9% of tests fall into the [0, 0.05) interval, 3.8% fall into the [0.05, 0.10) interval, and 3.7% fall into the [0.10, 0.15) interval. As the stars indicate (note their nonstandard interpretation, tailored to our context), the aforementioned results are all significantly different from 5%, the proportion arising in any interval of length 0.05 under the uniform benchmark. The 4.7% of p-values falling into the [0.15, 0.20) interval is not significantly different from the uniform benchmark. The pattern confirms what we observe from the kernel-density plot and suggests that misspecification and removal in the publication process are important forces in determining the shape of the p-curve.

TABLE 1.—REGRESSION RESULTS FOR PROPORTION OF SIGNIFICANT p-VALUES

Variable	Full sample	RCT balance, pure	Other tests, pure
Proportion $\hat{\pi}_j$ of p-values in subintervals			
$j = [0, 0.05)$	0.098***	0.046	0.091***
	(0.006)	(0.009)	(0.006)
$j = [0.05, 0.10)$	0.038***	0.025***	0.039***
	(0.002)	(0.005)	(0.003)
$j = [0.10, 0.15)$	0.037***	0.033***	0.039***
	(0.003)	(0.006)	(0.004)
$j = [0.15, 0.20)$	0.046	0.054	0.050
	(0.004)	(0.009)	(0.007)
Sum of three coefficients in removal region [0, 0.15)			
$\hat{\pi}_r$	0.173***	0.104***	0.169**
	(0.007)	(0.012)	(0.009)
Observation counts			
Unique sniff tests	28,837	2,953	16,063
Clusters	855	85	605

Columns show estimates from IFWLS regressions for select subsamples. See table A8 in the online appendix for the full set of regression results corresponding to every panel in figure 3. Estimates are based on more observations than kernel density plots: in addition to those for which we glean exact p-values, we include observations specifying p-value intervals. We use a stacked regression that allows each coefficient $\hat{\pi}_j$ to be estimated on the largest subsample I_j for which $\mathbf{1}_{p_i \in j}$ can be determined for any realization of $p_i \in [0, 1]$. The number of observations in each column differs slightly from that reported in figure 2 because some tables do not use conventional significance reporting thresholds. In the row for aggregate results over removal region [0, 0.15), entries equal the sum of the coefficients above in the same column: $\hat{\pi}_r = \hat{\pi}_{[0,0.05)} + \hat{\pi}_{[0.05,0.10)} + \hat{\pi}_{[0.10,0.15)}$. Standard errors reported in parentheses below results are clustered at the article level. The interpretation of stars is nonstandard here: rather than indicating a significant difference from 0, they indicate a significant difference from the uniform benchmark (with the proportion equaling the length of interval j) in a two-tailed test at the *10% level, **5% level, and ***1% level.

The next column provides regression results for the pure subsample of RCT balance tests. We expect little misspecification in this subsample, so absent removal, the p-curve should resemble the uniform benchmark, with the percentage of p-values significant in each interval equaling the size of the interval. Instead, we see that the $[0, 0.05)$, $[0.05, 0.10)$, and $[0.10, 0.15)$ intervals contain, respectively, 4.6%, 2.5%, and 3.3% of observations. The latter two percentages are significantly different from the 5% uniform benchmark at the 1% level. The [0.15, 0.20) interval, by contrast, contains 5.4% of p-values, which is slightly greater than the 5% uniform benchmark but not statistically significantly so.

The last column provides results for the pure subsample of other tests. Unlike pure RCT balance tests, for which we expect little misspecification, for the "mixed bag" of other tests, we do not know the extent of misspecification or the power of the tests employed to detect it. The results for the pure subsample of other tests are quite similar to for the full sample (perhaps unsurprisingly because the subsample constitutes the majority of the full sample). Both have one result that stands in sharp contrast to those for the pure subsample of RCT balance tests, namely, that the proportion of p-values in $[0, 0.05)$ is significantly greater than the uniform benchmark in the first and last columns but significantly less than the uniform benchmark in the middle column. Evidently, a substantial proportion of studies are failing sniff tests because of misspecification, enough to

TABLE 2.—ESTIMATES OF BOUNDS ON REMOVAL

Sample	Misspecification assumption	Lower bound $\underline{\rho}_r = \frac{r - \breve{\pi}_r}{r}$	Upper bound $\bar{\rho}_r = \frac{r - \breve{\pi}_r + \bar{\mu}_r \breve{\pi}_r}{r(1 - \breve{\pi}_r)}$
RCT balance, pure	Unknown misspecification ($\bar{\mu}_r = 1$ possible)	0.308***	1.116
		(0.077)	(0.014)
RCT balance, pure	No misspecification ($\bar{\mu}_r = 0$)	0.308***	0.344***
		(0.077)	(0.082)
Other tests, pure	Unknown misspecification ($\bar{\mu}_r = 1$ possible)	−0.128	1.204
		(0.057)	(0.012)
Other tests, pure	No misspecification ($\bar{\mu}_r = 0$)	−0.128	−0.154***
		(0.057)	(0.070)

Estimates of bounds on removal rate from proposition 1. Standard errors reported in parentheses below results computed using the delta method and clustered at the article level. The interpretation of stars is nonstandard here. For the lower-bound and upper-bound columns, stars indicate significantly greater than 0 and significantly less than 1, respectively, in a one-tailed test at the *10% level, **5% level, and ***1% level.

swamp the removal process that would tend to reduce the proportion of the most significant *p*-values.

Confirming the impression from a visual inspection of the kernel-density estimates, the regression results provide significant evidence of missing mass relative to the uniform benchmark—potential evidence for removal by the publication process—below the 0.15 threshold. Above this threshold, in the [0.15, 0.20) interval, there is no evidence of a departure from the uniform benchmark and thus no evidence of removal. Based on this evidence, we will take the removal region to be [0, 0.15) and thus $r = 0.15$ in the computations of bounds in the next subsection.[5]

C. *Estimates of Nonparametric Bounds*

The formulas provided in propositions 1 and 2 for bounds on the rates of removal and misspecification depend on just two variables, the threshold r dividing the removal from the no-removal region and the proportion $\breve{\pi}_r$ of *p*-values in the removal region. As discussed in the previous subsection, we set $r = 0.15$. Table 1 provides estimates $\hat{\pi}_r$ to input into the formulas. Standard errors for the output of these nonlinear formulas can be computed using the delta method.

Table 2 reports estimates of the bounds on the removal rate ρ_r provided by the formulas in proposition 1. The first row estimates bounds for the pure subsample of RCT balance tests without any assumptions on the extent of misspecification. The estimated lower bound is 30.8%. The estimate of the formula for the upper bound, 111.6%, exceeds the 100% mathematical limit on a rate, so it is uninformative. The next row provides estimates that bring in additional information on misspecification, assuming the rate is negligible for the pure RCT balance test subsample. Although the additional misspecification information does not help refine the lower bound, it does help refine the upper bound, yielding an estimate of 34.4%. Having sandwiched the removal rate between fairly narrow bounds, we can conclude—nonparametrically, solely from the missing mass in the removal region—that about a third of sniff tests were removed by the publication process from the pure sample of RCT balance tests assuming negligible misspecification in this subsample.[6]

The estimates of bounds on the removal rate for the pure subsample of other tests presented in the bottom row of the table turn out to be uninformative. Without assumptions on the misspecification rate, the estimated bounds are wider than the mathematical limits on a rate, [0%, 100%]. Under the assumption of no misspecification in the subsample, the estimated upper bound becomes negative—an impossibility—indicating that the assumption is violated for this subsample. There must be at least some misspecification in the pure subsample of other tests because the proportion of *p*-values in [0, 0.15) in the postremoval subsample at 17% exceeds the 15% uniform benchmark, an excess which can only be explained by misspecification.[7]

Table 3 reports estimates of the bounds on the misspecification rate μ_r provided by the formulas in proposition 2. The first row estimates bounds for the pure subsample of RCT balance tests without any assumptions on the extent of removal. The bounds are uninformative because they are wider than the [0%, 100%] mathematical limits on a rate. The next row brings in additional information from the estimated the lower bound on the removal rate for that subsample. Although the additional information helps

[5] Elliott et al. (2022) also adopt the [0, 0.15) removal region for their empirical applications.

[6] Andrews and Kasy (2019) provide a method for estimating publication removal using information on coefficients and standard errors in published articles. Their method requires true effects and standard errors to be independent, which appears to be violated by clusters of observations obvious in the scatterplot for our pure sample of RCT balance tests. We instead sought to apply their methods to subsamples for individual journals. Among journals exhibiting little clustering in the scatterplot of coefficients versus standard errors, the *American Economic Review* (*AER*) supplies the most usable observations (sniff tests reporting both coefficients and standard errors). Modifying their web application posted on the GitHub site maxkasy.github.io to allow a 15% threshold for the removal region and translating the resulting estimate into a removal rate generates a 17% removal rate for the *AER*. Applying our methods to the *AER* subsample, we bound the removal rate between 23% and 26%, which comes close to enveloping the 17% estimate using the Andrews and Kasy (2019) method.

[7] In the online appendix, we tighten the lower bound on removal in the absence of misspecification information for the pure sample of other tests to 9.3% by exploiting the additional information that comes from partitioning the removal region into subintervals. The bound draws on Theorem 1 of Elliott et al. (2022), which states that the *p*-curve is nonincreasing under general conditions.

TABLE 3.—ESTIMATES OF BOUNDS ON MISSPECIFICATION

Sample	Removal assumption	Lower bound $\underline{\mu}_r = \frac{\breve{\pi}_r - r + r(1-\breve{\pi}_r)\underline{\rho}_r}{(1-r)\breve{\pi}_r}$	Upper bound $\bar{\mu}_r = 1$
RCT balance, pure	Unknown or no removal ($\underline{\rho}_r = 0$)	−0.524	1.000
		(0.190)	
RCT balance, pure	Lower bound estimated for this sample ($\underline{\rho}_r = 0.308$)	−0.054	1.000
		(0.014)	
Other tests, pure	Unknown or no removal ($\underline{\rho}_r = 0$)	0.134***	1.000
		(0.053)	
Other tests, pure	Lower bound estimated for RCT balance, pure sample ($\underline{\rho}_r = 0.308$)	0.401***	1.000
		(0.076)	

Estimates of bounds on misspecification rate from proposition 2. Standard errors reported in parentheses below results computed using the delta method and clustered at the article level. Because trivial upper bound $\bar{\mu}_r = 1$ is posited, not estimated, standard errors omitted for those entries. The interpretation of starsin the lower-bound column is nonstandard here, indicating significantly greater than 0 in a one-tailed test at the *10% level, **5% level, and ***1% level.

narrow the bounds on the misspecification rate, the bounds remain uninformative. To understand why, recall that proposition 2 places no upper bound on the misspecification rate; as explained in the Theory section, even a *p*-curve that looks uniform can hide a large mass of slightly misspecified studies. The proposition provides a lower bound leveraging excess mass in the removal region relative to the uniform benchmark. However, the pure subsample of RCT balance tests exhibits missing mass in this region rather than an excess, making it impossible to rule out that the subsample exhibits no misspecification (indeed this is a maintained assumption in computing a set of removal bounds above).

The bounds on misspecification for the pure subsample of other tests are informative. If the removal rate is unknown (or known to be 0), the estimated lower bound is 13.4%.[8] If one assumes the removal rate is the same as for the pure subsample of RCT balance tests, which is bounded below by 30.8%, the estimate of $\underline{\mu}_r$ rises to a substantial 40.1%.

D. *Text Analysis of Author Claims*

Authors have incentives to attribute unfavorable sniff tests to random bad luck. Such claims are difficult to dispute on an individual basis. To investigate whether authors, in the aggregate, tend to over- or underattribute unfavorable sniff tests to bad luck, we combine our nonparametric bounds on the latent proportion of misspecified studies with hand-collected data on authors' qualitative characterization of their own sniff-test results.

Our research assistants read the discussions of sniff tests in the articles and rated the authors' confidence in the test result according to a rubric involving four categories: "strong claim," "weak claim," "admit rejected," and "no claim." 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 not significant. 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 while 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 pretreatment 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. "Admit rejected" includes cases when authors freely acknowledge that the significant *p*-value indicates rejection of the sniff test and a potential problem for their study. When the authors do not discuss the specific test statistic, we classify it as "no claim."

In our pure subsample of RCT balance tests, restricting attention to *p*-values in the [0, 0.15) removal region, authors make strong claims in 78% of cases (employing the same inverse frequency weighting used for all our empirical results). These claims are not unjustified under our assumption that misspecification in RCT is negligible.

In our pure sample of other tests, 52% of authors with *p*-values in the removal region make strong claims.[9] Whether these claims can be justified by our estimated misspecification rates depends on some assumptions. If we assume nothing about the removal rate, allowing for the possibility of no removal, then according to table 3 the misspecification rate in the preremoval population is bounded below by approximately 14%, implying that the proportion of well-specified studies in the preremoval population is bounded above by 86%. If no studies are removed, this 86% upper bound on

[8] In the online appendix, we tighten the lower bound on misspecification in the absence of removal information for the pure sample of other tests to 22.3% by exploiting the additional information that comes from partitioning the removal region into subintervals. The bound draws on a lemma proved in the appendix that the proportion of well-specified studies is bounded by the largest rectangle that can be inscribed under the *p*-curve.

[9] See table A9 in the online appendix for a full set of results on authors' claims.

well-specified studies is inherited by the postremoval population. The fraction of authors making strong claims is well within this bound. Instead, assume that studies with significant other tests are removed at the same rate as studies with significant RCT balance tests. Furthermore, assume that the removal rate is homogeneous across p-values in the removal region (as the estimates in table 1 for the pure sample of RCT balance tests suggest) and independent of whether studies are truly misspecified (which is unobservable to the publication process). Then the relevant lower bound on misspecification from table 3 is 40%, implying that the proportion of well-specified studies is bounded above by 60%. The fraction of authors making strong claims continues to be inside the bound, but not by much, suggesting that such claims might merit reader scrutiny.

VI. Conclusion

This paper analyzed a hand-collected sample of nearly 30,000 sniff tests from scores of economics journals. Our most detailed analysis focused on what we call "pure" samples in which p-value improvement techniques were unlikely to have been used. We further divided the sample into subsamples, one for RCT balance tests and one for other tests.

A visual inspection revealed stark differences in the the p-curves across the two subsamples. For the pure subsample of RCT balance tests, the p-curve is missing mass relative to the uniform benchmark, evidence of removal by the publication process (whether due to p-hacking or relegation to the metaphorical "file drawer"). The pure subsample of other tests has extra mass piled on low p-values, evidence of a high rate of misspecification in the associated published studies.

Despite being unable to observe the full sample of studies prior to removal by the publication process, we were able to construct nonparametric bounds on the latent rates of removal and misspecification among these studies using only information on the proportion of significant p-values in published studies. Estimates of these bounds vary depending on the strength of the assumptions behind them. Under the plausible assumption that there is negligible misspecification in RCT balance tests, arising only in the rare instance of a failed randomization, we estimated that the publication process removed about a third of significant RCT balance tests. Under the plausible assumption that RCT balance tests and other tests experienced similar removal rates, we estimated that more than 40% of significant sniff tests represented actual misspecification, not simply an unlucky sniff-test draw for a well-specified study.

REFERENCES

Andrews, Isaiah, "Valid Two-Step Identification-Robust Confidence Sets for GMM," this REVIEW 100:2 (2018), 337–348.

Andrews, Isaiah, Matthew Gentzkow, and Jesse M. Shapiro, "On the Informativeness of Descriptive Statistics for Structural Estimates," *Econometrica* 88:6 (2020), 2231–2258. 10.3982/ECTA16768

Andrews, Isaiah, and Maximilian Kasy, "Identification of and Correction for Publication Bias," *American Economic Review* 109:8 (2019), 2766–2794. 10.1257/aer.20180310

Ashenfelter, Orley, Colm Harmon, and Hessel Oosterbeek, "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias," *Labour Economics* 6:4 (1999), 453–470. 10.1016/S0927-5371(99)00041-X

Athey, Susan, and Guido W. Imbens, "The Econometrics of Randomized Experiments" (pp. 73–140), in Esther Duflo and Abhijit Banerjee (eds.), *Handbook of Economic Field Experiments*, vol. 1 (Amsterdam: North-Holland, 2017). 10.1016/bs.hefe.2016.10.003

Bilinski, Alyssa, and Laura A. Hatfield, "Nothing to See Here? Non-Inferiority Approaches to Parallel Trends and Other Model Assumptions," arXiv:1805.03273 (2020).

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting Event Study Designs: Robust and Efficient Estimation," SSRN working paper 2826228 (2022).

Brodeur, Abel, Nikolai Cook, and Anthony Heyes, "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review* 110:11 (2020), 3634–3660. 10.1257/aer.20190687

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg, "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics* 8:1 (2016), 1–32. 10.1257/app.20150044

Bruhn, Miriam, and David McKenzie, "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics* 1:4 (2009), 200–232. 10.1257/app.1.4.200

Card, David, and Alan B. Krueger, "Time-Series Minimum-Wage Studies: A Meta-Analysis," *American Economic Review Papers and Proceedings* 85:2 (1995), 238–243.

Christensen, Garret. S., and Edward Miguel, "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature* 56:3 (2018), 920–980. 10.1257/jel.20171350

DeLong, J. B., and Kevin Lang, "Are All Economic Hypotheses False?" *Journal of Political Economy* 100:6 (1992), 1257–1272. 10.1086/261860

Dette, Holger, and Martin Schumann, "Testing for Equivalence of Pretrends in Difference-in-Differences Estimation," *Journal of Business & Economic Statistics* 42:3 (2024), 1289–1301. 10.1080/07350015.2024.2308121

Doucouliagos, Chris, "Publication Bias in the Economic Freedom and Economic Growth Literature," *Journal of Economic Surveys* 19:3 (2005), 367–387. 10.1111/j.0950-0804.2005.00252.x

Elliott, Graham, Nikolay Kudrin, and Kaspar Wüthrich, "Detecting P-Hacking," *Econometrica* 90:2 (2022), 887–906. 10.3982/ECTA18583

Freyaldenhoven, Simon, Christian Hansen, and Jesse M. Shapiro, "Pre-Event Trends in the Panel Event-Study Design," *American Economic Review* 109:9 (2019), 3307–3338. 10.1257/aer.20180609

Görg, Holger, and Eric Strobl, "Multinational Companies and Productivity Spillovers: A Meta-Analysis," *Economic Journal* 111:475 (2001), 723–739.

Havránek, Tomáš, "Measuring Intertemporal Substitution: The Importance of Method Choices and Selective Reporting," *Journal of the European Economic Association* 13:6 (2015), 1180–1204.

Hung, H. M. James, Robert T. O'Neill, Peter Bauer, and Karl Kohne, "The Behavior of the P-Value When the Alternative Hypothesis Is True," *Biometrics* 53:1 (1997), 11–22. 10.2307/2533093

Ioannidis, John P. A., "Inverse Publication Reporting Bias Favoring Null, Negative Results," *BMJ Evidence-Based Medicine* 29:1 (2024), 6–9. 10.1136/bmjebm-2023-112292

Ioannidis, John P. A., and Chris Doucouliagos, "What's to Know about the Credibility of Empirical Economics?" *Journal of Economic Surveys* 27:5 (2013), 997–1004. 10.1111/joes.12032

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos, "The Power of Bias in Economics Research," *Economic Journal* 127:605 (2017), F236–F265. 10.1111/ecoj.12461

Jann, Ben, "KDENS: Stata Module for Univariate Kernel Density Estimation," Statistical Software Component S456410, Department of Economics, Boston College (2005), available at: http://ideas.repec.org/c/boc/bocode/s456410.html.

Jones, M. C., "Simple Boundary Correction for Kernel Density Estimation," *Statistics and Computing* 3 (1993), 135–146. 10.1007/BF00147776

Kahn-Lang, Ariella, and Kevin Lang, "The Promise and Pitfalls of Differences-in-Differences: Reflections on *16 and Pregnant* and Other Applications," *Journal of Business & Economic Statistics* 38:3 (2019), 1–14.

Lehmann, E. L., and Joseph P. Romano, *Testing Statistical Hypotheses* (New York: Springer, 2005).

Morgan, Kari L., and Donald B. Rubin, "Rerandomization to Improve Covariate Balance in Experiments," *Annals of Statistics* 40 (2012), 1263–1282. 10.1214/12-AOS1008

Nelson, Jon P., "Estimating the Price Elasticity of Beer: Meta-Analysis of Data with Heterogeneity, Dependence, and Publication Bias," *Journal of Health Economics* 33 (2014), 180–187. 10.1016/j.jhealeco.2013.11.009

Rosenthal, Robert, "The 'File Drawer' Problem and Tolerance for Null Results," *Psychological Bulletin* 86:3 (1979), 638–641. 10.1037/0033-2909.86.3.638

Roth, Jonathan, "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends," *American Economic Review: Insights* 4:3 (2022), 305–322. 10.1257/aeri.20210236

Stanley, T. D., "Beyond Publication Bias," *Journal of Economic Surveys* 19:3 (2005), 309–345. 10.1111/j.0950-0804.2005.00250.x