Examining Selection Pressures in the Publication Process Through the Lens of Sniff Tests

Christopher M. Snyder
Dartmouth College and NBER

Ran Zhuo
University of Michigan

Abstract: The increasing demand for empirical rigor has led to the growing use of auxiliary tests (balance, pre-trends, over-identification, placebo, etc.) to help 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. We use these sniff tests—a sample of nearly 30,000 hand collected from scores of economics journals—as a lens to examine selection pressures in the publication process. We derive bounds under plausible nonparametric assumptions 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. For the subsample of balance tests in randomized controlled trials, we find that the publication process removed at least 30% of significant p-values. For the subsample of other tests, we find a that at least 40% of significant p-values indicated true misspecification. We use textual analysis to assess whether authors over-attribute significant sniff tests to bad luck.

JEL classification: C18, A14, B41

Acknowledgments: 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, 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.

1. Introduction

The increasing demand for rigor in empirical economics has led to the increasing use of auxiliary tests (balance, parallel trends, over-identification, placebo, etc.) 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 to not only require sniff tests to be reported but 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, Cook, and Heyes 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 893 articles in 59 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 $[0,1]$ by construction. 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 (re-randomization, 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 non-uniform, 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 misspecification. One of our contributions in this paper is the derivation of such bounds under plausible nonparametric assumptions. Initially, we only impose the minimal assumption that misspecification tends to pile mass on low p-values while 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 suggest that 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.

2. 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. While their study examines 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 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, Jaravel, and Spiess (2022) propose new tests for pre-trends in event studies. Andrews, Gentzkow, and Shapiro (2020) propose a measure relating the performance of sniff tests to their informativeness on the robustness of main results. Freyaldenhoven, Hansen, and Shapiro (2019) propose methods to estimate causal effects in event studies when pre-trends are present in the outcomes. Instead of the null of no misspecification, Bilinski and Hatfield (2020) and Dette and Schumann (2022) 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 pre-analysis 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, 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. Our methods allow us to pool observations across a range of topics, as do Brodeur *et al.* (2016) and Ioannidis, Stanley, and Doucouliagos (2017).

Our approach shares some similarities with recent advances in this literature. Elliott, Kudrin, and Wüthrich (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

[1] The medical literature is just beginning to recognize what Ioannidis (2023) dubs inverse publication bias, arising in clinical safety trials and other contexts in his taxonomy.

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 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.

3. 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," etc. (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*, *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 double checked by supervising research assistants. Our final sample includes 29,776 sniff-test observations reported in 1,369 different tables from 893 articles published in 59 journals.[2] 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 sniff tests in our sample grows at a 48% average annual rate.

Figure 2 breaks down our sample into subsamples by methodology. It is important to distinguish among methodologies since the distribution of p-values can differ systematically across them. A key breakdown is between tests of balance in randomized controlled trials (RCTs) and

[2] Our raw dataset includes both one- and two-sided tests. The difficulties of analyzing publication bias in the context of one-sided tests raised by DeLong and Lang (1992) led us to restrict the final sample to two-sided tests. We dropped a further 764 observations that either were not structured as well-defined hypothesis tests or did not provide sufficient information to glean either an exact p-value or an interval for it. We arrive at our final sample by dropping two additional observations reporting p-values exceeding 1.

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 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 is comprised by 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 re-randomization (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 and Rubin 2012).

Balance-improvement techniques shift some of the mass of low p-values to higher values, creating a departure from the benchmark uniform distribution. 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 re-randomization by taking a conservative approach to constructing the pure subsample, only including items for which re-randomization 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 re-randomization in the article but we cannot definitively rule out re-randomization because authors had access to the baseline data to re-randomize 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. Re-randomization and stratification are irrelevant for these observations since 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 which use sniff tests to assess the validity of their causal-identification strategies, removing studies that fail the pre-tests 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 are likely to be greater with other tests than RCT balance tests, presuming 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, 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.

4. 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.

4.1. 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, i.e.,

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

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

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

Let $\mathscr{F}(p,\theta)$ denote the 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,

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

For a textbook proof and technical conditions required for (3) to hold, see, e.g., Lehmann and Romano (2005, Lemma 3.3.1). Notice that the uniform benchmark in (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, i.e.,

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

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.[3]

[3] To prove (4), let $\beta(\alpha,\theta)$ denote the power of a sniff test with significance level α against al-

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 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. Since result (3) holds for all cdfs over which $F_0(p)$ is integrated, $F_0(p)$ inherits the result, implying $F_0(p) = p$. Result (4) holds for all cdfs over which $F_1(p)$ is integrated, implying $F_1(p) \geq p$. Together, results (3) and (4) imply $F(p) \geq p$.

ternative $\theta \in \Theta_1$, i.e., $\beta(\alpha,\theta) = \Pr(Z_\theta \in R(\alpha))$. 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 (2), and the third step from definition of power. The final step holds for unbiased sniff tests. Exchanging the variable p for α yields (4).

4.2. 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 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 post-removal populations of studies. Let m denote the mass of p-values in the pre-removal population, π_r the proportion in the removal region of pre-removal population, and π_n the proportion in the no-removal region of the pre-removal population. Analogues of the variables just defined written with breves (a mnemonic for change, in this case resulting from removal) refer to their counterparts for the post-removal population. Thus, $\breve{m}$ denotes the mass of sniff tests in the post-removal 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. Since 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 (plims) to which consistent finite-sample estimators converge. Our dataset is a sample of the post-removal 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 since 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 since ρ_r depends on unobserved properties of the pre-removal 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 post-removal are accounted for, in every region and in the population as a whole.

The identity

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

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 post-removal 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 pre-removal mass of p-values is preserved post-removal, implying $\breve{m} = m = 1$ by (9). The proportion of p-values in the removal region (also the mass since the unit mass is preserved) differs by the amount of the shift: $s_r = \pi_r - \breve{\pi}_r$ by (10). Substituting the preceding equality along with $d_r = 0$ into (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 (10), further substituting $\breve{m} = 1 - d_r$ from (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 post-removal population. Substituting this value of d_r

along with $s_r = 0$ into (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 (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 post-removal 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 post-removal 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). Since there are no other forms of removal besides shifting and deletion, we have established that, under the null of no misspecification, (11) provides a lower bound $\underline{\rho}_r$ on the removal rate and (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 (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 post-removal mass. In the presence of misspecification, the upper bound in (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 post-removal distribution. As this construction suggests, the upper bound on the removal rate in the presence of possible misspecification is the trivial one, $\bar{\rho}_r = 1$.

4.3. Nonparametric Bounds on Misspecification

The pre-removal 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 pre-removal population. We will focus on bounding not μ but μ_r, the misspecification rate restricted to the removal region. The reason 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 (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 (13). Noting that $F_0(r) = r$ as shown in Section 4.1 and that $F_1(r) \leq 1$ since F_1 is an average of cdfs, substituting these expressions into (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 $\mu F_1(r) = \pi_r - (1-\mu)r \geq (\pi_r - r)/(1-r)$, where the equality follows from rearranging (13) and the inequality from (15). Substituting this inequality into the numerator of (14) yields

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

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

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

The right-hand side of (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 pre-removal 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 (13)–(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 post-removal population, there may have been a greater departure pre-removal, requiring even more misspecification to explain. The lower bound $\underline{\mu}_r$ in (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.

4.4. 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 we, 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 (12) on removal in the absence of misspecification can be recovered from Proposition 1 by substituting $\bar{\mu}_r = 0$ into (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 (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 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 pre-removal sample in bounding rates of removal and misspecification. While 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 post-removal rather than the pre-removal population. Owing to space constraints, we provide a comprehensive analysis of these alternatives in the online appendix.

5. 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 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.

5.1. 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 since 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 (re-randomization, 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 pre-matched 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 post-matched sample. Since the post-matched sample is used for estimation, the publication process likely exerts little pressure to remove significant sniff tests for pre-matched 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 post-match. We see quite the opposite of the pre-match 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.

5.2. 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, 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.[4] For example, a table reporting

[4] 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.

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.[5] 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:

$$\breve{\pi}_j = E_{t \in T_j}(E(\mathbf{1}_{p_i \in j} \,|\, i \in t)), \tag{21}$$

The conditional expectation in (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).

A consistent estimator $\hat{\pi}_j$ of the population proportion $\breve{\pi}_j$ can be obtained by replacing the expectations in (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

[5] 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.

convenient, equivalent expression for $\hat{\pi}_j$ by introducing inverse frequency weights

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

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

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

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 $i \in I_j$, i.e.,

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

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. Since 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 misspecification and removal in the publication process are important forces in determining the shape of the p-curve.

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 since 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 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.[6]

5.3. 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 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. While 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

[6] Elliot, Kudrin, and Wüthrich (2022) also adopt the $[0, 0.15)$ removal region for their empirical applications.

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.[7]

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 since the proportion of p-values in $[0, 0.15)$ in the post-removal subsample at 17% exceeds the 15% uniform benchmark,

[7] 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 the extent of heterogeneity in the scatterplot for our pure sample of RCT balance tests. We apply their methods instead to individual-journal subsamples with less evidence of heterogeneity. Among journals with little evidence of heterogeneity, the two supplying the most usable observations (sniff tests reporting standard errors as well as coefficients) are the *Journal of Economic Behavior & Organization* (*JEBO*) and *American Economic Review* (*AER*). Modifying their web application posted on the GitHub site "Estimating publication bias using meta-studies—Maximilian Kasy" (maxkasy.github.io) to allow a 15% threshold for the removal region and translating the resulting estimate into a removal rate generates a 70% removal rate for *JEBO* and 17% for the *AER*. Our methods applied to these individual journal subsamples bound the removal rate between 55% and 59% for *JEBO* and between 28% and 31% for the *AER*. This comparison suggests that while the alternative methods do not produce identical results, they are directionally similar.

an excess which can only be explained by misspecification.[8]

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 since 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. While the additional information helps narrow the bounds on the misspecification rate, the bounds remain uninformative. To understand why, recall 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 13.6%.[9] If one assumes

[8] 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 9.2% by exploiting the additional information that comes from partitioning the removal region into subintervals. The bound draws on Theorem 1 of Elliott, Kudrin, and Wüthrich (2022), which states that the p-curve is nonincreasing under general conditions.

[9] 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 22.4% by exploiting the additional information that comes from partitioning the removal region into subintervals. The bound draws on

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 40.2%.

5.4. 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 under-attribute 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 pre-treatment

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.

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.[10] 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 pre-removal population is bounded below by approximately 14%, implying that the proportion of well-specified studies in the pre-removal population is bounded above by 86%. If no studies are removed, this 86% upper bound on well-specified studies is inherited by the post-removal 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

[10] See Table A9 in the online appendix for a full set of results on authors claims.

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.

6. Conclusion

This paper analyzed a hand-collected sample of nearly 30,000 sniff tests from 893 articles in 59 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, only arising 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 over 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,” *Review of Economics and Statistics* 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.

Andrews, Isaiah and Maximilian Kasy, “Identification of and Correction for Publication Bias,” *American Economic Review* 109:8 (2019), 2766–2794.

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.

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).

Bilinski, Alyssa and Laura A. Hatfield, “Nothing to See Here? Non-inferiority Approaches to Parallel Trends and Other Model Assumptions,” arXiv preprint 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, Mathias Lé, Marc Sangnier, and Yanos Zylberberg, "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics* 8:1 (2016), 1–32.

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.

Bruhn, Miriam and David McKenzie, "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics* 1:4 (2019), 200–232.

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.

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

Dette, Holger, and Martin Schumann, "Testing for Equivalence of Pre-trends in Difference-in-Differences Estimation," Maastricht University working paper (2022).

Doucouliagos, Chris, "Publication Bias in the Economic Freedom and Economic Growth Literature," *Journal of Economic Surveys* 19:3 (2005), 367–387.

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

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.

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

Havránek, Tomáš. (2015). "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.

Ioannidis, John P. A., "Inverse Publication Reporting Bias Favoring Null, Negative Results," *BMJ Evidence-Based Medicine* forthcoming (2023).

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.

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

Jann, Ben, "KDENS: Stata module for univariate kernel density estimation," Statistical Software Component S456410, Department of Economics, Boston College (2005). 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.

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. (2005) *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.

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.

Rosenthal, Robert, “The ‘File Drawer’ Problem and Tolerance for Null Results,” *Psychological Bulletin* 86:3 (1979), 638–641.

Roth, Jonathan, “Pretest with Caution: Event-Study Estimates After Testing for Parallel Trends,” *American Economic Review: Insights* 4:3 (2022), 305–322.

Sellke, Thomas, M. J. Bayarri, and James O. Berger, “Calibration of p Values for Testing Precise Null Hypotheses,” *American Statistician* 55:1 (2001), 62–71.

Stanley, T. D., “Beyond Publication Bias,” *Journal of Economic Surveys* 19:3 (2005), 309–345.

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

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			
• $j = [0, 0.05)$	0.099^{***} (0.006)	0.046 (0.009)	0.091^{***} (0.006)
• $j = [0.05, 0.10)$	0.038^{***} (0.002)	0.025^{***} (0.005)	0.039^{***} (0.003)
• $j = [0.10, 0.15)$	0.037^{***} (0.003)	0.033^{***} (0.006)	0.039^{***} (0.004)
• $j = [0.15, 0.20)$	0.047 (0.004)	0.054 (0.009)	0.050 (0.007)
Aggregate proportion $\hat{\pi}_r$ over removal region $[0, 0.15)$	0.173^{***} (0.007)	0.103^{***} (0.012)	0.170^{**} (0.009)
Observation counts			
• Unique sniff tests	28,832	2,953	16,063
• Clusters	857	85	606

Notes: 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 row for aggregate results over removal region $[0, 0.15)$, entries equal the sum of the coefficients above in 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 non-standard 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 *ten-percent level, **five-percent level, ***one-percent level.

Table 2: Estimates of Bounds on Removal

	Lower bound	Upper bound
Sample, $\bar{\mu}_r$ assumption	$\underline{\rho}_r = \frac{r-\breve{\pi}_r}{r}$	$\bar{\rho}_r = \frac{r-\breve{\pi}_r+\bar{\mu}_r\breve{\pi}_r}{r(1-\breve{\pi}_r)}$
RCT balance, pure		
• Unknown misspecification ($\bar{\mu}_r = 1$ possible)	0.308***	1.116
	(0.078)	(0.015)
• No misspecification ($\bar{\mu}_r = 0$)	0.308***	0.344***
	(0.078)	(0.082)
Other tests, pure		
• Unknown misspecification ($\bar{\mu}_r = 1$ possible)	−0.131	1.204
	(0.057)	(0.012)
• No misspecification ($\bar{\mu}_r = 0$)	−0.131	−0.157***
	(0.057)	(0.070)

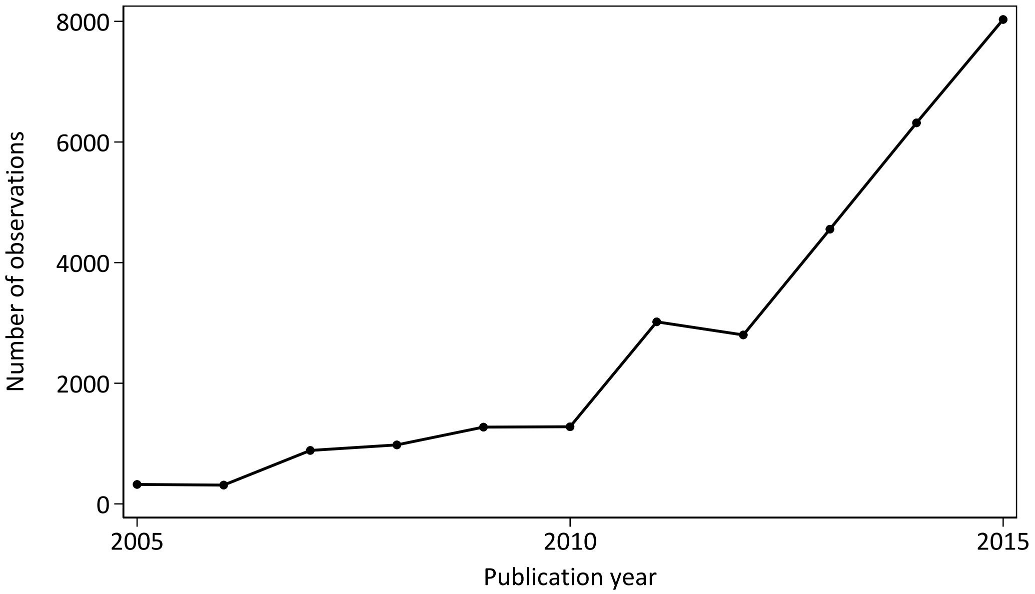
Notes: Estimates of bounds on removal rate from Proposition 1. Standard errors reported in parentheses below results computed using the delta method and clustered at the article level. The interpretation of stars is non-standard here. For the lower-bound and upper-bound columns, stars indicate significantly greater than 0 and significantly less than 1, respectively, in a one-tailed test at the *ten-percent level, **five-percent level, ***one-percent level.

Table 3: Estimates of Bounds on Misspecification

	Lower bound	Upper bound
Sample, $\underline{\rho}_r$ assumption	$\underline{\mu}_r = \frac{\breve{\pi}_r-r+r(1-\breve{\pi}_r)\underline{\rho}_r}{(1-r)\breve{\pi}_r}$	$\bar{\mu}_r = 1$
RCT balance, pure		
• Unknown or no removal ($\underline{\rho}_r = 0$)	−0.524	1.000
	(0.191)	
• Lower bound estimated for this sample ($\underline{\rho}_r = 0.308$)	−0.054	1.000
	(0.014)	
Other tests, pure		
• Unknown or no removal ($\underline{\rho}_r = 0$)	0.136***	1.000
	(0.052)	
• Lower bound estimated for RCT balance, pure sample ($\underline{\rho}_r = 0.308$)	0.402***	1.000
	(0.075)	

Notes: Estimates of bounds on misspecification rate from Proposition 2. Since trivial upper bound $\bar{\mu}_r = 1$ is posited, not estimated, standard errors omitted for those entries. See previous table for additional notes.

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: Subsample Breakdown by Methodology

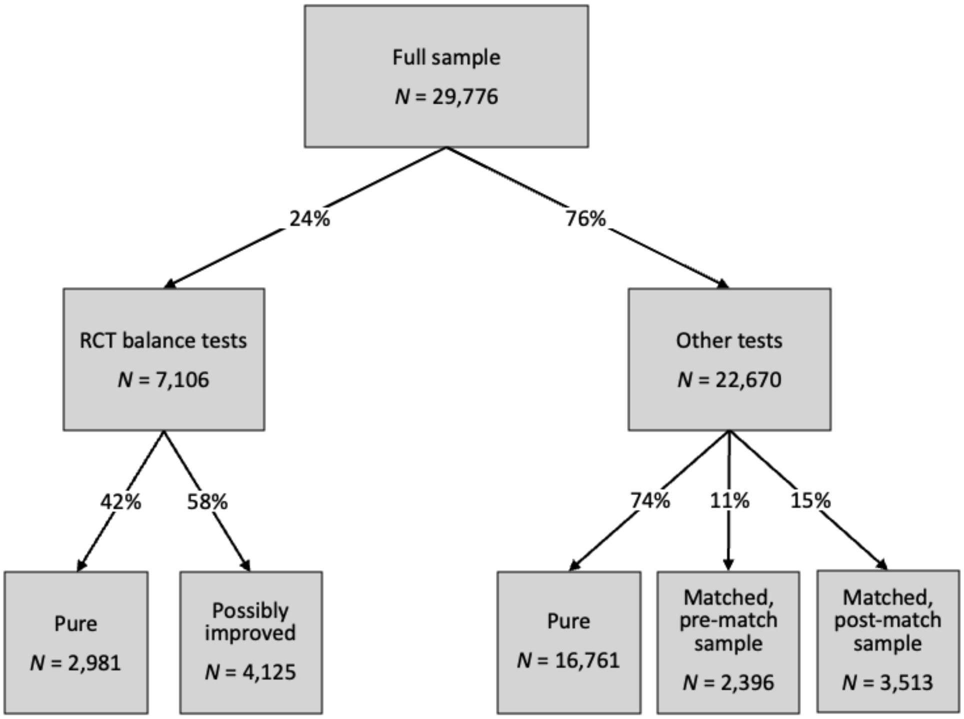
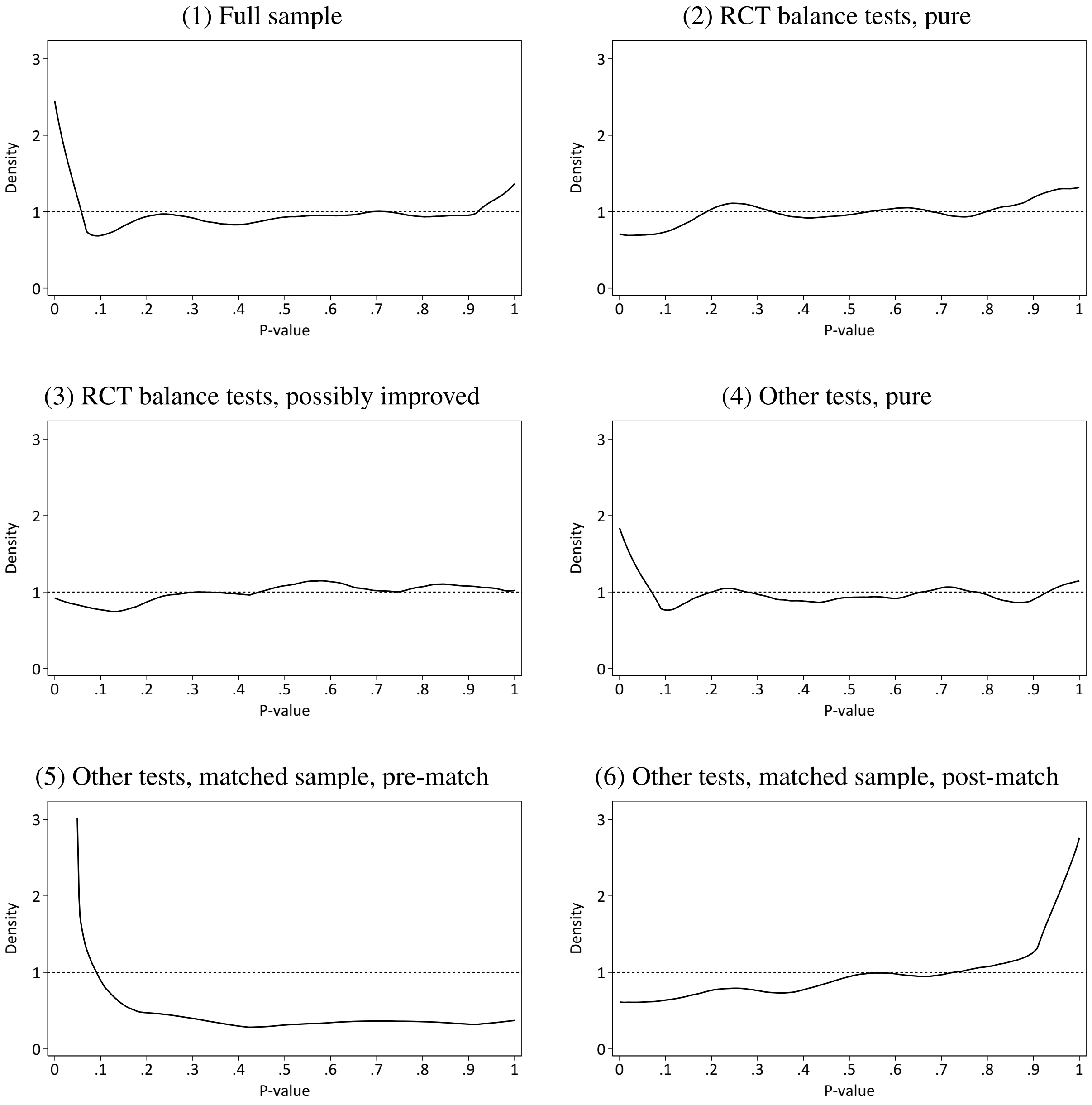


Figure 3: Kernel Density Estimates of P-curves


(1) Full sample

(2) RCT balance tests, pure

(3) RCT balance tests, possibly improved

(4) Other tests, pure

(5) Other tests, matched sample, pre-match

(6) Other tests, matched sample, post-match

Notes: 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.

Online Appendixes

Examining Selection Pressures in the Publication Process Through the Lens of Sniff Tests

Christopher M. Snyder
Dartmouth College and NBER

Ran Zhuo
University of Michigan

This document contains a series of online appendixes supplementing the article. The appendixes include the following content.

Appendix A1 provides details behind the construction of the sample.

Appendix A2 provides bounds for the case in which the removal region is treated as a single interval. This is the appendix containing proofs of the propositions stated in the main text.

Appendix A3 provides bounds for alternative definitions of removal and misspecification rates than adopted in the text, for example defining rates with respect to the post-removal rather than pre-removal population or defining rates over the entire unit interval rather than just in the removal region.

Appendix A4 generalizes the analysis to allow the removal region to be partitioned into subintervals with potentially different removal rates.

Appendix A5 provides details behind the inverse frequency weighting used in our regressions.

Appendix A6 provides full regression tables for results mentioned in the main text.

A1. Details on Sample Construction

Search Phrases for Sniff Tests

We construct the initial pool of articles potentially containing sniff tests from searches of keywords in Elsevier's online database, ScienceDirect, We include any article that was published in the list of Elsevier journals in Table A1 and that contains any of the following phrases:

- "balance test"
- "balancing test"
- "baseline characteristics"
- "falsification checks"
- "falsification test"
- "placebo regression"
- "placebo test"
- "randomization checks"
- "randomization test"
- "validation checks"
- "validation test"
- ("compare" OR "comparison") AND "at baseline"
- ("compare" OR "comparison") AND "treatment group" AND "control group" AND ("before treatment" OR ("prior" AND "treatment")).

We similarly construct the initial pool of articles potentially containing sniff tests from the list of top-five, general-interest journals in Table A1 by searches of the JSTOR database for any of the listed phrases. We performed the same keyword search on Google Scholar to obtain the initial pool of articles for years 2013–15 for the *Quarterly Journal of Economics* and 2015 for the *Review of Economic Studies*, journal volumes which were not available through JSTOR during our initial data collection in 2015.

Sample Journals

Table A1 lists the journals in economics and affiliated fields contained in the sample, grouped by five-year impact factor.

Table A1: Journals in Sample by Five-Year Impact Factor

Impact Factor ≥ 4			Impact Factor $\in [2,4)$			Impact Factor < 2		
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	15.9	*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	7.8	*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	10.2	*J. Ec. Behavior & Org.*	1.5	5.1
			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

Notes: Listing of journals constituting sample, divided into three categories by five-year impact factor. 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 the third category 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% of the sample; together, they constitute less than 1.5%. 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.

A2. Nonparametric Bounds with a Single Removal Region

Overview

This appendix first proves the two propositions stated in the main text, which provide nonparametric bounds when the unit interval of p-values is partitioned into just two regions, the removal region $[0, r)$, and the no-removal region $[r, 1]$. The appendix then states and proves propositions on when those bounds are sharp. Finally, we provide an alternative to equation (19) for an upper bound on removal and discuss the circumstances under which one or the other bound is tighter.

Proof of Proposition 1

Lower bound (18) can be established with the following series of inequalities:

$$\rho_r \equiv \frac{d_r + s_r}{\pi_r} = \frac{\pi_r - \breve{m}\breve{\pi}_r}{\pi_r} \geq \frac{r - \breve{m}\breve{\pi}_r}{r} \geq \frac{r - \breve{\pi}_r}{r}. \tag{A1}$$

The first step repeats the definition in (8). The second step follows from admissibility constraint (10). The third step follows from

$$\pi_r = F(r) \geq r, \tag{A2}$$

where the equality follows from the definition of the cdf and the inequality is shown in Section 4.1 to hold by (4). The fourth step follows from $\breve{m} \leq 1$, which in turn follows from the admissibility constraint (9).

We next turn to establishing upper bound (19). Before doing so, we will establish two preliminary facts. Section 4.1 argued that the asymptotic uniformity of the p-value distribution captured in equation (3) implies $F_0(r) = r$. Substituting the preceding equality into (13) yields

$$\pi_r = (1-\mu)r + \mu F_1(r) = (1-\mu)r + \pi_r \mu_r \leq r + \pi_r \bar{\mu}_r. \tag{A3}$$

The second equality follows from substituting for $\mu F_1(r)$ from (14). The last inequality follows from $\mu \geq 0$ and $\mu_r \leq \bar{\mu}_r$. Rearranging (A3) yields our first preliminary fact:

$$\pi_r \leq \frac{r}{1 - \bar{\mu}_r}. \tag{A4}$$

The second preliminary fact comes from combining the admissibility conditions (9) and (10) and rearranging, yielding

$$\pi_r = \breve{\pi}_r + (1 - \breve{\pi}_r)d_r + s_r. \tag{A5}$$

Proceeding to establish the upper bound, substituting (A5) into (8) yields

$$\rho_r = \frac{d_r + s_r}{\breve{\pi}_r + (1 - \breve{\pi}_r)d_r + s_r}. \tag{A6}$$

Substituting (A5) into (A4) yields

$$\breve{\pi}_r + (1 - \breve{\pi}_r)d_r + s_r \leq \frac{r}{1 - \bar{\mu}_r}. \tag{A7}$$

We can bound ρ_r above by maximizing (A6) subject to (A7) and nonnegativity constraints $d_r \geq 0$ and $s_r \geq 0$. By (A6), we will be assured that ρ_r does not lie above the maximum of the resulting value function, so we will have our upper bound.

Objective (A6) is increasing in d_r and s_r. Hence, constraint (A7) binds at the optimum. Treating (A7) as an equality, solving for s_r, substituting this value into the right-hand side of objective (A6), and rearranging yields

$$1-\frac{\breve{\pi}_r(1-d_r)(1-\bar{\mu}_r)}{r}. \tag{A8}$$

This rewritten objective is increasing in d_r and independent of s_r. Hence, it is maximized subject to (A7) by setting $s_r^* = 0$ and

$$d_r^* = \frac{1}{1-\breve{\pi}_r}\left(\frac{r}{1-\bar{\mu}_r} - \breve{\pi}_r\right), \tag{A9}$$

the value of d_r for which (A7) holds with equality setting $s_r = 0$. (That the corner involves no shifting, only deletion, is expected from the discussion in Section 4.2, which argued that attributing removal to deletion rather than shifting generates the greatest removal rate consistent with observed properties of the post-removal sample.) Substituting d_r^* and s_r^* into (A6) and rearranging yields the upper bound in (19). □

Proof of Proposition 2

We showed in (A3) that

$$\pi_r = (1-\mu)r + \mu F_1(r). \tag{A10}$$

We will draw two implications from (A10). First, a simple rearrangement shows

$$\mu F_1(r) = \pi_r - (1-\mu)r. \tag{A11}$$

The second implication follows from the fact that, since F_1 is an average over cdfs, it inherits the property $F_1(r) \leq 1$. Substitituing the preceding inequality into (A10) yields $\pi_r \leq (1-\mu)r + \mu$, which upon rearranging yields

$$\mu \geq \frac{\pi_r - r}{1-r}. \tag{A12}$$

Leveraging those two implications allows us to derive a lower bound on the misspecification rate:

$$\mu_r \equiv \frac{\mu F_1(r)}{\pi_r} = \frac{\pi_r - (1-\mu)r}{\pi_r} \geq \frac{\pi_r - r}{(1-r)\pi_r}. \tag{A13}$$

The first step repeats the definition of μ_r in (14). The second step follows from substituting from (A11). The last step follows by substituting from inequalty (A12) and rearranging. We proceed by solving the problem of choosing the removal configuration $\{d_r, s_r\}$ minimizing the rightmost expression in (A13) subject to admissibility constraints. By (A13), we will be assured that μ_r does not lie below the minimum of the resulting value function, so we will have our lower bound.

One can verify that the rightmost expression in (A13) is increasing in π_r. The rest of the terms in the expression are exogenous. Hence, the removal configuration minimizing the rightmost expression in (A13) subject to constraints will be the same as that minimizing π_r subject to constraints. Substituting from for $\breve{m}$ from admissibility constraint (9) into admissibility constraint

(10) and rearranging yields

$$\pi_r = \breve{\pi}_r + (1-\breve{\pi}_r)d_r + s_r. \tag{A14}$$

We will minimize the right-hand side of (A14) subject to nonnegativity constraints $d_r \geq 0$ and $s_r \geq 0$ as well as a constraint incorporating the researcher's knowledge of the bound on removal rate:

$$\rho_r \equiv \frac{d_r + s_r}{\pi_r} = \frac{d_r + s_r}{\breve{\pi}_r + (1-\breve{\pi}_r)d_r + s_r} \geq \underline{\rho}_r, \tag{A15}$$

where the first equality repeats the definition of ρ_r from (8), the second equality follows from substituting from (A14), and the third reflects the knowledge that the removal rate is no less than some threshold $\underline{\rho}_r$. Rearranging the last inequality in (A15) yields a useful form for the removal constraint

$$[1-(1-\breve{\pi}_r)\underline{\rho}_r]d_r + (1-\underline{\rho}_r)s_r \geq \underline{\rho}_r\breve{\pi}_r. \tag{A16}$$

Minimizing the right-hand side of (A14) subject to (A16) and nonnegativity constraints is a linear program. It can be solved by checking the value of the objective at the two vertices of the constraint set. One vertex sets $s_r = 0$ and sets d_r to the value satisfying (A16) with equality:

$$d_r = \frac{\underline{\rho}_r\breve{\pi}_r}{1-(1-\breve{\pi}_r)\underline{\rho}_r}. \tag{A17}$$

The other vertex sets $d_r = 0$ and sets s_r to the value satisfying (A16) with equality:

$$s_r = \frac{\underline{\rho}_r\breve{\pi}_r}{1-\underline{\rho}_r}. \tag{A18}$$

One can verify that the first vertex produces a lower solution and thus provides the mininum for the linear program. Substituting the solution into (A14), and substituting the resulting value of π_r into the original objective function given by the rightmost expression in (A13) yields the bound stated in equation (20). □

Sharpness of Bounds

Next, we state a result on the sharpness of one of the bounds just established.

Proposition 3. *Suppose there is no misspecification in the pre-removal sample. Then the bounds from Proposition 1, upon substituting $\bar{\mu}_r = 0$, are sharp.*

Proof. To prove the lower bound in (18) is sharp, consider the configuration of removals such that $d_r^* = 0$ and $s_r^* = r - \breve{\pi}_r$, which turns out to satisfy the combination of admissibility conditions in (A5) when $d_r = 0$. Substituting d_r^* and s_r^* into the removal rate as expressed in (A6) yields $\rho_r = (r-\breve{\pi}_r)/r$, the bound in (18). Demonstrating an admissible configuration of removals attaining the lower bound proves it is sharp.

To prove the upper bound in (19) is sharp, consider the configuration of removals such that $s_r = 0$ and $d_r^* = (r-\breve{\pi}_r)/(1-\breve{\pi}_r)$, which turns out to satisfy (A5) when $s_r = 0$. Substituting d_r^* and s_r^* into (A6) yields $\rho_r = (r-\breve{\pi}_r)/r(1-\breve{\pi}_r)$, the bound in (19) when $\bar{\mu}_r = 0$. Demonstrating an admissible configuration of removals attaining the upper bound proves it is sharp. □

Alternative Bounds

The next proposition provides an alternative upper bound on the removal rate to that stated in Proposition 1, which is tighter under some conditions.

Proposition 4. *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 above the upper bound*

$$\bar{\bar{\rho}}_r = \frac{r - \breve{\pi}_r + (1-r)\bar{\mu}_r}{(1-\breve{\pi}_r)[r + (1-r)\bar{\mu}_r]}. \tag{A19}$$

This bound is tighter than (19) *if and only if*

$$\bar{\mu}_r > \frac{1-2r}{1-r} \tag{A20}$$

and weaker if the reverse inequality holds.

Proof. We can obtain an upper bound on ρ_r by choosing d_r and s_r to maximize ρ_r as expressed in (A6) subject to nonnegativity constraints $d_r \geq 0$ and $s_r \geq 0$ and the combination of the remaining admissibility conditions as expressed in (A5). Now

$$\pi_r = (1-\mu_r)r + \mu_r F_1(r) \leq r + (1-r)\mu_r \leq r + (1-r)\bar{\mu}_r, \tag{A21}$$

where the equality follows from (A10), the first inequality from $F_1(r) \leq 1$, and the last inequality from the fact that $\bar{\mu}_r$ is an upper bound on μ_r. Combining (A5) with (A21) yields a weaker constraint

$$\breve{\pi}_r + (1-\breve{\pi}_r)d_r + s_r \leq r + (1-r)\bar{\mu}_r. \tag{A22}$$

Following the logic used to solve similar problems in this appendix, the solution to this weaker problem of maximizing (A6) subject to nonnegativity and (A22) involves $s_r^* = 0$ and the value d_r^* solving (A22) treated as an equality after substituting $s_r = 0$, i.e.,

$$d_r^* = \frac{r - \breve{\pi}_r + (1-r)\bar{\mu}_r}{1-\breve{\pi}_r}. \tag{A23}$$

Substituting d_r^* and s_r^* into (A6) yields the bound in (A19).

To verify (A20), the difference between (19) and (A19) can be written, after rearranging,

$$\frac{\breve{\pi}_r \bar{\mu}_r[(1-r)(1-\bar{\mu}_r) - r]}{r(1-\breve{\pi}_r)[r + (1-r)\bar{\mu}_r]}. \tag{A24}$$

All the factors are definitively positive except the one in square brackets, which is positive if and only if (A20) holds, in which case (19) exceeds (A20) and so (A20) is a tighter bound. □

A3. Alternatively Defined Rates

For consistency, the main text focused on the the removal region of the pre-removal population in bounding rates of removal and misspecification. While 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 post-removal rather than the pre-removal population.

Table A2 provides a taxonomy of the rates analyzed in the main text and the additional alternatives analyzed in this appendix. Each box provides the notation for the alternative rate. It also cross-references the tables in which the relevant bound formulas and bound estimates from our dataset are reported. Subsequent sections of this appendix derive the formulas for the rates in Table A2 and lower and upper bounds on those rates thus defined.

Table A2: Taxonomy of Alternatively Defined Rates

		Population			
		Pre-removal		Post-removal	
Region	Removal	ρ_r Table 2	μ_r Table 3	$\breve{\rho}_r$ No table	$\breve{\mu}_r$ Tables A5 and 3
	Unit	ρ Table A3	μ Table A4	$\breve{\rho}$ No table	$\breve{\mu}$ Tables A6 and A7

Notes: Since there is no removal from the post-removal population, the removal rates $\breve{\rho}_r$ and $\breve{\rho}$ are not well defined and thus are not reported in tables. Several entries have two tables of reported results covering various subcases to be defined.

Removal Rate from Entire Pre-removal Population

Since the mass of the pre-removal population is normalized to $m = 1$, the removal rate from the entire pre-removal population equals the total mass removed:

$$\rho = d_r + s_r. \tag{A25}$$

Rearranging (10) yields $d_r + s_r = \pi_r - \breve{m}\breve{\pi}_r \geq r - \breve{\pi}_r$, where the last inequality follows from (A2). This provides a lower bound on the removal rate: $\underline{\rho} = r - \breve{\pi}_r$.

Next turn to deriving an upper bound on (A25) when we, the researchers, know $\mu_r \leq \bar{\mu}_r$. Similar logic as in the proof of Proposition 1 can be used to show that a bound can be obtained by choosing $d_r \geq 0$ and $s_r \geq 0$ to maxizing removal rate (A25) subject to (A7). The solution is the same here as in that proof: d_r^* given by (A9) and $s_r^* = 0$. Substituting those values into objective

(A25) yields upper bound

$$\bar{\rho} = \frac{r - \breve{\pi}_r(1-\bar{\mu}_r)}{(1-\breve{\pi}_r)(1-\bar{\mu}_r)}. \tag{A26}$$

The next table summarizes the rate and bound formulas just derived and reports estimates of the bounds in our sniff-test sample.

Table A3: Bounds on Removal Rate ρ in Entire Pre-removal Population

Sample, $\bar{\mu}_r$ assumption	Lower bound $\underline{\rho} = r - \breve{\pi}_r$	Upper bound $\bar{\rho} = \frac{r-\breve{\pi}_r(1-\bar{\mu}_r)}{(1-\breve{\pi}_r)(1-\bar{\mu}_r)}$
RCT balance, pure		
• Unknown misspecification ($\bar{\mu}_r = 1$ possible)	0.046***	∞
	(0.012)	
• No misspecification ($\bar{\mu}_r = 0$)	0.046***	0.052***
	(0.012)	(0.012)
Other tests, pure		
• Unknown misspecification ($\bar{\mu}_r = 1$ possible)	−0.019	∞
	(0.009)	
• No misspecification ($\bar{\mu}_r = 0$)	−0.019	−0.024***
	(0.009)	(0.011)

Notes: Notes: Estimates of bounds on removal rate, formulas for which are provided in column headings. The negative estimate for $\bar{\rho}$ is inconsistent with the requirement that the removal rate be nonnegative, indicating that the assumption of no misspecification is invalid. See Table 2 for additional notes.

Misspecification Rate in Entire Pre-removal Population

The misspecification rate in the entire pre-removal population is simply μ. The argument in Section 4.3 that $\bar{\mu}_r = 1$ is a tight upper bound on μ_r carries over here to imply that $\bar{\mu} = 1$ is a tight upper bound on μ.

We are left to derive a lower bound, denoted $\underline{\mu}$ when we, the researchers, know $\rho_r \geq \underline{\rho}_r$. By (A12), a lower bound on μ is given by

$$\frac{\pi_r - r}{1-r} = \frac{\breve{\pi}_r + (1-\breve{\pi}_r)d_r + s_r - r}{1-r}, \tag{A27}$$

where the equality follows from substititing for π from (10). To arrive at lower bound $\underline{\mu}$, we can choose $d_r \geq 0$ and $s_r \geq 0$ to minimize the right-hand side of (A27) subject to $\rho_r = (d_r + s_r)/\pi_r \geq \underline{\rho}_r$, which was shown in the proof of Proposition 2 to be equivalent to (A16). The upshot is a linear program, which one can verify by testing the vertices to have the same solution as in that proof:

d_r^* given by (A17) and $s_r^* = 0$. Substituting that solution into the objective given by the right-hand side of (A27) yields lower bound

$$\underline{\mu} = \frac{\breve{\pi}_r - r + r(1-\breve{\pi}_r)\underline{\rho}_r}{(1-r)[1-(1-\breve{\pi}_r)\underline{\rho}_r]}. \tag{A28}$$

The next table summarizes the rate and bound formulas just derived and reports estimates of the bounds in our sniff-test sample.

Table A4: Bounds on Misspecification Rate μ in Entire Pre-removal Population

	Lower bound	Upper bound
Sample, $\underline{\rho}_r$ assumption	$\underline{\mu} = \frac{\breve{\pi}_r - r + r(1-\breve{\pi}_r)\underline{\rho}_r}{(1-r)[1-(1-\breve{\pi})\underline{\rho}_r]}$	$\bar{\mu}_r = 1$
RCT balance, pure		
• Unknown or no removal ($\underline{\rho}_r = 0$)	−0.054 (0.014)	1.000
• Lower bound estimated for this sample ($\underline{\rho}_r = 0.308$)	−0.008 (0.002)	1.000
Other tests, pure		
• Unknown or no removal ($\underline{\rho}_r = 0$)	0.023** (0.010)	1.000
• Lower bound estimated for RCT balance, pure sample ($\underline{\rho}_r = 0.308$)	0.092*** (0.026)	1.000

Notes: Estimates of bounds on misspecification rate, formulas for which are provided in column headings. See Table 3 for additional notes.

Misspecification Rate in Removal Region of Post-removal Population Under Weak Assumptions on Misspecification Among Removed

The misspecification rate in the removal region of the post-removal population is given by

$$\breve{\mu}_r = \frac{\mu_r \pi_r - \mu_{ds}(d_r + s_r)}{\breve{m}\breve{\pi}_r}, \tag{A29}$$

where μ_{ds} is the misspecification rate among removed observations. Presumably, the misspecification rate among removed observations is at least as great as the unconditional misspecification rate μ_r in the removal region pre-removal even if in the worst case removal is at random. We will place no restrictions on μ_{ds} other than the natural constraint that a rate cannot entail more than complete misspecification: $\mu_{ds} \leq 1$.

The argument in Section 4.3 that $\bar{\mu}_r = 1$ is a tight upper bound on μ_r carries over here to imply that $\bar{\breve{\mu}} = 1$ is a tight upper bound on $\breve{\mu}$.

We are left to derive a lower bound, denoted $\underline{\breve{\mu}}$ when we, the researchers, know $\rho_r \geq \underline{\rho}_r$. We have

$$\breve{\mu}_r \geq \frac{\mu_r \pi_r - d_r - s_r}{\breve{m}\breve{\pi}_r} \tag{A30}$$

$$\geq \frac{\pi_r - r - (1-r)(d_r + s_r)}{(1-r)\breve{m}\breve{\pi}_r} \tag{A31}$$

$$= \frac{\breve{\pi}_r - r}{(1-r)\breve{\pi}_r} + \frac{rs_r}{(1-r)(1-d_r)\breve{\pi}_r}, \tag{A32}$$

where (A30) follows from substituting $\mu_{ds} \leq 1$, (A31) follows from substituting $\mu_r \geq (\pi_r - r)/(1-r)\pi_r$ from (A13) and rearranging, and (A32) follows from further rearranging. Equation (A32) can be minimized by setting $s_r^* = 0$ without violating the constraint entailed by knowledge of $\rho_r \geq \underline{\rho}_r$. Substituting $s_r^* = 0$ into (A32) yields lower bound

$$\underline{\breve{\mu}} = \frac{\breve{\pi}_r - r}{(1-r)\breve{\pi}_r}. \tag{A33}$$

The next table summarizes the rate and bound formulas just derived and reports estimates of the bounds in our sniff-test sample. Notice that the weak assumptions on misspecification rate μ_{ds} among removed observations preclude our leveraging knowledge of the removal rate, so $\underline{\rho}_r$ is not a factor in this table.

Table A5: Bounds on Misspecification Rate $\breve{\mu}_r$ in Removal Region of Post-removal Population Under Weak Assumptions on Misspecification Among Removed

Sample	Lower bound $\underline{\mu} = \frac{\breve{\pi}_r - r}{(1-r)\breve{\pi}_r}$	Upper bound $\bar{\mu}_r = 1$
RCT balance, pure	−0.524 (0.014)	1.000
Other tests, pure	0.136*** (0.052)	1.000

Notes: Estimates of bounds on misspecification rate, formulas for which are provided in column headings. The weak assumptions on misspecification rate μ_{ds} among removed observations preclude our leveraging knowledge of the removal rate, so $\underline{\rho}_r$ does is not a factor in this table. See Table 3 for additional notes.

Misspecification Rate in Entire Post-removal Population Under Weak Assumptions on Misspecification Among Removed

The misspecification rate in the entire post-removal population is given by

$$\breve{\mu} = \frac{\mu - \mu_{ds} d_r}{\breve{m}}. \tag{A34}$$

We again place no restrictions on the misspecification rate among removed observations, μ_{ds}, other than $\mu_{ds} \leq 1$.

The argument in Section 4.3 that $\bar{\mu}_r = 1$ is a tight upper bound on μ_r carries over here to imply that $\bar{\breve{\mu}} = 1$ is a tight upper bound on $\breve{\mu}$.

We are left to derive a lower bound, denoted $\underline{\breve{\mu}}$ when we, the researchers, know $\rho_r \geq \underline{\rho}_r$. We have

$$\begin{aligned}
\breve{\mu} &\geq \frac{\mu - d_r}{\breve{m}} && \text{(A35)}\\
&= \frac{\mu - d_r}{1 - d_r} && \text{(A36)}\\
&\geq \frac{\pi_r - r - (1-r)d_r}{(1-r)(1-d_r)} && \text{(A37)}\\
&= \frac{\breve{\pi}_r + (1-\breve{\pi}_r)d_r + s_r - r - (1-r)d_r}{(1-r)(1-d_r)}, && \text{(A38)}
\end{aligned}$$

The constraint $\rho_r \geq \underline{\rho}_r$ was shown in the proof of Proposition 2 to be equivalent to (A16). We can thus compute a lower bound by minimizing (A38) subject to (A16). The constraint turns out to bind. Solving the constraint treated as an equality for s_r, substituting into the objective, and differentiating the result with respect to d_r yields a negative derivative, implying that the optimum involves a high value of d_r^* and low value of s_r^*. Substituting $s_r^* = 0$ into the objective yields lower bound

$$\underline{\breve{\mu}} = \frac{\breve{\pi} - r}{1 - r}. \tag{A39}$$

The next table summarizes the rate and bound formulas just derived and reports estimates of the bounds in our sniff-test sample. Notice again that the weak assumptions on misspecification rate μ_{ds} among removed observations preclude our leveraging knowledge of the removal rate, so as in Table A5, $\underline{\rho}_r$ is not a factor in this table either.

Table A6: Bounds on Misspecification Rate $\breve{\mu}$ in Entire Post-removal Population Under Weak Assumptions on Misspecification Among Removed

Sample	Lower bound $\underline{\breve{\mu}} = \frac{\breve{\pi}_r - r}{1-r}$	Upper bound $\breve{\bar{\mu}}_r = 1$
RCT balance, pure	-0.054 (0.014)	1.000
Other tests, pure	0.023^{**} (0.010)	1.000

Notes: Notes from Table A5 apply.

Misspecification Rate in Removal Region of Post-removal Population Assuming Random Removal

Equation (A29) shows the misspecification rate in the removal region of the post-removal population. If removal is random, implying that the removal rate is independent of whether studies are well specified or misspecified and indepenent of the p-value within the removal region, then $\mu_{sd} = \mu_r$. Substituting into (A29) yields

$$\breve{\mu}_r = \frac{\mu_r(\pi_r - d_r - s_r)}{\breve{m}\breve{\pi}_r} = \mu_r, \tag{A40}$$

where the second equality follows from substituting for the denominator from (10). This is the identical formula for the misspecification rate studied in the main text for the pre-removal population. See the main text for bounds on this misspecification rate and Table 3 for bounds estimates in our sniff-test sample.

Misspecification Rate in Entire Post-removal Population Assuming Random Removal

Equation (A34) shows the misspecification rate in the entire post-removal population. As argued in the previous section, $\mu_{sd} = \mu_r$ if removal is random. Substituting into (A34) yields

$$\begin{aligned}
\breve{\mu} &= \frac{\mu - \mu_r d_r}{\breve{m}} && \text{(A41)}\\
&= \frac{\mu[\pi_r - F_1(r)d_r]}{\pi_r(1-d_r)} && \text{(A42)}\\
&\geq \frac{\mu(\pi_r - d_r)}{\pi_r(1-d_r)} && \text{(A43)}\\
&\geq \frac{(\pi_r - r)(\pi_r - d_r)}{(1-r)\pi_r(1-d_r)} && \text{(A44)}
\end{aligned}$$

$$\geq \frac{(\breve{\pi}_r - r)(\breve{\pi}_r - d_r)}{(1-r)\breve{\pi}_r(1-d_r)} \tag{A45}$$

Equation (A42) follows from substituting for μ from (9) and for μ_r from (14). Condition (A43) follows from $F_1(r) \leq 1$. Condition (A44) follows from (A12). To see (A45), differentiating (A44) with respect to π_r yields a positive derivative. But $\breve{\pi}_r \leq \pi_r$ since $\breve{\pi}_r$ reflects removed mass. Thus, the substitution of $\breve{\pi}_r$ for π_r in moving from (A44) to (A45) reduces the expression.

A lower bound can be obtained by minimizing (A45) subject to the constraint that $\rho_r \geq \underline{\rho}_r$, which was shown previously to be equivalent to (A16). Since s_r^* does not appear in (A45), the solution sets $s_r^* = 0$ and d_r^* such that (A16) holds with equality after substituting $s_r^* = 0$, i.e., the value of d_r^* satisfying (A17). Substituting from (A17) into objective (A45) and yields lower bound

$$\underline{\breve{\mu}} = \frac{\breve{\pi}_r - r + r(1-\breve{\pi}_r)\underline{\rho}_r}{1-r}. \tag{A46}$$

The next table summarizes the rate and bound formulas just derived and reports estimates of the bounds in our sniff-test sample.

Table A7: Bounds on Misspecification Rate $\breve{\mu}$ in Entire Post-removal Population Assuming Random Removal

Sample, $\underline{\rho}_r$ assumption	Lower bound $\underline{\mu} = \frac{\breve{\pi}_r - r + r(1-\breve{\pi}_r)\underline{\rho}_r}{1-r}$	Upper bound $\bar{\mu}_r = 1$
RCT balance, pure		
• Unknown or no removal ($\underline{\rho}_r = 0$)	−0.054 (0.014)	1.000
• Lower bound estimated for this sample ($\underline{\rho}_r = 0.308$)	−0.006 (0.001)	1.000
Other tests, pure		
• Unknown or no removal ($\underline{\rho}_r = 0$)	0.023** (0.010)	1.000
• Lower bound estimated for RCT balance, pure sample ($\underline{\rho}_r = 0.308$)	0.068*** (0.015)	1.000

Notes: Estimates of bounds on misspecification rate, formulas for which are provided in column headings. See Table 3 for additional notes.

A4. Nonparametric Bounds with Multiple Removal Regions

Overview

In this appendix, we generalize the analysis from the previous appendix to allow for an arbitrary number of removal regions with potentially differing rates of shifting and deletion. As we will explain, the move to multiple removal regions does not tighten many of the bounds derived with a single removal region since many of them were shown to be sharp. We are able to tighten two of the bounds as stated in Propositions 5 and 6 below. The appendix concludes with an empirical estimation of the improved bounds using our sniff-test sample.

Setup

To pursue this generalization, some additional notation is in order. Consider a partition $0 = a_0 < a_1 < \ldots < a_J \leq a_{J+1} = 1$ of the unit interval. Suppose there is possible removal of p-values in the first J subintervals $[a_{j-1}, a_j)$, $j = 1, \ldots, J$, but no removal in the last subinterval $[a_J, 1]$, which is allowed to be an empty set when $a_J = 1$.

As before, the breve accent is used to distinguish between the pre- and post-removal populations of p-values. Denote the proportion of observations falling into interval j in the pre-removal population by π_j and in the post-removal population by $\breve{\pi}_j$. Since the unit interval is partitioned by subintervals $j = 1, \ldots, J+1$, we have $\sum_{j=1}^{J+1} \pi_j = \sum_{j=1}^{J+1} \breve{\pi}_j = 1$. Denote the total mass in the pre-removal population by m and in the post-removal population by $\breve{m}$. Normalize $m = 1$. Suppose the pre-removal p-value distribution is continuous with probability density function (pdf) $f(p)$.

Let d_j denote the mass of p-values deleted from interval j and $s_{j,j'}$ denote the mass shifted out of j into interval j'. Assume $d_{J+1} = 0$ and $s_{j,j'} = 0$ for all $j' < j$. These assumptions ensure that p-hacking and other mechanisms only shift mass up, not down, and that no mass is removed from the rightmost interval $J+1$, consistent with its definition as falling outside the removal region.

The propositions below are concerned with bounds on the rate of removal and misspecification in the removal region $[0, a_J)$ pre-removal. To establish the bounds, we will show that there exists no admissible configuration of removals $\{d_j, s_{j,j'} \,|\, j = 1, \ldots, J; j' > j\}$ for which the relevant rate lies outside the bound. As discussed in the text, to be admissible, the configuration of removals must be nonnegative,

$$d_j \geq 0 \text{ for all } j = 1, \ldots, J \tag{A47}$$

$$s_{j,j'} \geq 0 \text{ for all } j = 1, \ldots, J \text{ and } j' > j, \tag{A48}$$

must ensure that overall mass in both samples is accounted for,

$$\breve{m} = m - \sum_{j=1}^{J} d_j = 1 - \sum_{j=1}^{J} d_j, \tag{A49}$$

and must ensure that mass in each interval j is accounted for,

$$\breve{m}\breve{\pi}_j = \pi_j - d_j - \sum_{j'>j} s_{j,j'} + \sum_{j'<j} s_{j',j} \tag{A50}$$

for all $j = 1,\ldots,J+1$.

The notation for the more general model here can be connected back to the notation in the main text. The threshold for removal, is given by $r \equiv a_J$. The proportion of observations in the removal region and no removal region, respectively, in the pre-removal population are given by $\pi_r \equiv \sum_{j=1}^{J} \pi_j$ and $\pi_n \equiv \pi_{J+1}$. Analogously, the proportion of observations in the removal region and no removal region, respectively, in the post-removal population are given by $\breve{\pi}_r \equiv \sum_{j=1}^{J} \breve{\pi}_j$ and $\breve{\pi}_n \equiv \breve{\pi}_{J+1}$. Total deletions amount to $d_r \equiv \sum_{j=1}^{J} d_j$, and total shifts out of the removal region amount to $s_r \equiv \sum_{j=1}^{J} s_{j,J+1}$.

The rest of this appendix maintains the same assumptions required for the background results in Section 4.1 to hold, including that sniff tests have nested rejection regions, are unbiased, and so forth. For brevity, we omit the list of those maintained assumptions from the statement of the propositions; the reader is referred to that subsection for a list and discussion of the maintained assumptions.

Inherited Bounds from a Single Removal Region

Having set up the more general model, we proceed to derive bounds on removal and misspecification with multiple removal regions. We seek results in a similar vein as those with a single removal region: bounding the removal rate given the knowledge we, the researchers, have about the misspecification rate and vice versa. With multiple removal regions, there is some ambiguity in how the misspecification rate should be constrained, whether the constraint is on the aggregate misspecification rate μ_r over the entire removal region or whether separate constraints μ_j should be specified for each of the removal intervals $j = 1,\ldots,J$. We resolve this ambiguity by assuming the knowledge we, the researchers, have takes on one of the extremes, either knowing that there is no misspecification or knowing nothing about the misspecification rate and thus leaving it unconstrained. The same ambiguity arises when considering we, the researchers, knowledge of the removal rate when bounding misspecification. We will resolve that ambiguity similarly by assuming that either there is no removal or the removal rate is unconstrained.

Start with the derivation of bounds on the removal rate. Assume first that there is no misspecification in the removal region, $\mu_r = 0$, implying that there is no misspecification in any of the subintervals of the removal region. The bounds derived with a single removal region continue to apply with multiple removal regions. In principle, the bounds could be tighter with in the latter case. However, Proposition 3 shows they are sharp with a single removal region, so must be sharp with multiple removal regions.

Now assume that the researcher does not know the misspecification rate. We argued that the tight upper bound on removal is $\bar{\rho}_r = 1$ with a single removal region. The argument relied on a construction with increasingly extreme misspecification in which π_r is increasingly large, requiring an increasing removal rate to arrive at the observed post-removal distribution. The same argument implies that $\bar{\rho}_r = 1$ with multiple removal regions as well. This leaves the open question of whether the lower bound on removal $\underline{\rho}_r$ with no information on misspecification can be improved by moving to multiple removal regions. We prove in the next section that it can.

We next turn to deriving bounds on the misspecification rate μ_r given either that the researcher knows $\rho_r = 0$ or the researcher is uninformed about ρ_r. We argued in the text that the upper bound on the misspecification rate with a single removal region is the trivial $\bar{\mu}_r = 1$, which continues to be the upper bound with multiple removal regions. This leaves the open queston of whether the lower bound on misspecification $\underline{\mu}_r$ can be improved by moving to multiple removal regions. Again, we prove in the next section that it can.

Improving Bounds Over Single Removal Region

We will analyze the open questions raised in the previous section in reverse order, starting a proof that the lower bound $\underline{\mu}_r$ on misspecification can be improved with multiple removal regions. We argued in the text that the bound when the researcher knows the removal rate is $\rho_r = 0$ is the same as when the researcher does not know the removal rate because $\rho_r = 0$ is the "worst case" for generating misspecification. The next proposition in this appendix is phrased as bounding the misspecification rate when the researcher lacks information on the removal rate, but it could be equivalently phrases as bounding the misspecification rate when $\rho_r = 0$. The bound formula is complex, reflecting an iterative procedure we call "backward ironing" which we use to calculate the shifts that complement deletions needed to minimize the misspecification rate to obtain the lower bound. Despite the complex formula, the procedure can be described in intuitive terms and illustrated with a diagram.

Before stating the last proposition, we provide a lemma stating a useful intermediate result. To streamline the statement of the next lemma and proof of the subsequent proposition, let $\underline{f} \equiv \min\{f(p) \,|\, p \in [0,1]\}$.

Lemma 1. *Consider a continuous random variable with pdf $f(p)$ and support in $[0,1]$. We can write $f(p) = (1-\mu) + \mu g(p)$ for some well-defined density function $g(p)$ and weight $\mu \in [0,1]$ if and only if $\mu \in [1-\underline{f}, 1]$.*

Proof. See Pounds and Morris (2003) for a proof of the "if" direction. To prove the "only if" direction, suppose we write $f(p) = (1-\mu) + \mu g(p)$ for some $\mu < 1-\underline{f}$. Then there exists p' in the support of f such that $f(p') = \underline{f} < 1-\mu$. Thus, $(1-\mu) + \mu g(p') < 1-\mu$, implying $g(p') < 0$, implying g cannot be a well-defined density. □

The lemma is relevant for the next proposition, implying that the mass of well-specified studies cannot exceed the largest rectangle that can be inscribed under the pdf f of p-values in the pre-removal sample. Residual mass between the top of this inscribed rectangle and f must reflect misspecification; those observations must have something other than a uniform distribution that would have arisen absent misspecification.

Proposition 5. *Suppose we, the researchers, are uninformed about the removal rate ρ_r. There exists no admissible configuration of removals $\{d_j, s_{j,j'} \,|\, j = 1,\ldots,J; j' > j\}$ such that the misspecification rate*

$$\mu_r = \frac{\mu F_1(a_J)}{\sum_{j=1}^{J} \pi_j}, \tag{A51}$$

falls below lower bound

$$\underline{\mu}_r = \frac{1 - \breve{\pi}_{J+1} - a_J + (1-a_J)\sum_{j=1}^{J} d_j^*}{(1-a_J)\left(1 - \breve{\pi}_{J+1} + \sum_{j=1}^{J} d_j^*\right)}, \tag{A52}$$

where

$$d_j^* \equiv \max\left[0, \delta_j + \sum_{j'=1}^{j-1} s_{j',j}^* - \sum_{j'=j+1}^{J} s_{j,j'}^*\right], \tag{A53}$$

and

$$\delta_j \equiv \left(\frac{a_j - a_{j-1}}{1 - a_J}\right)\breve{\pi}_{J+1} - \breve{\pi}_j. \tag{A54}$$

The iterative procedure for deriving the optimal $s_{j',j}^*$ is embodied by the following recursive formula:

$$s_{j',j}^* = \max\left[0, \frac{1}{1+e_{j-1}}\left(\delta_{j'} - \delta_j + \sum_{k=1}^{j'-1} s_{k,j'}^*\right)\right], \tag{A55}$$

where

$$e_j = 1 + \min\left\{j' = 1, \ldots, j-1 \,|\, s_{j',j}^* > 0\right\} \tag{A56}$$

is the number of subintervals that were equalized with j in stage j ironing.

Proof. Before proceeding, we establish a preliminary inequality:

$$\sum_{j=1}^{J} \pi_j = F(a_J) \tag{A57}$$

$$= (1-\mu)F_0(a_J) + \mu F_1(a_J) \tag{A58}$$

$$= (1-\mu)a_J + \mu F_1(a_J) \tag{A59}$$

$$\leq \underline{f} a_J + \mu F_1(a_J) \tag{A60}$$

$$\leq \min_{j\in\{1,\ldots,J+1\}}\left(\frac{\pi_j}{a_j - a_{j-1}}\right) a_J + \mu F_1(a_J) \tag{A61}$$

Equation (A57) follows from (5), equation (A58) from (13), and equation (A59) from (3). Condition (A60) follows from Lemma 1. Condition (A61) follows because the lowest value $\underline{f}$ of the pdf of the pre-removal sample over $[0,1]$ cannot exceed the average density $\pi_j/(a_j - a_{j-1})$ in any subinterval.

Substituting for $\mu F_1(a_J)$ from (A57)–(A61) into the numerator of (A51) yields

$$\mu_r \geq 1 - \frac{a_J}{\sum_{j=1}^{J} \pi_j} \min_{j\in\{1,\ldots,J+1\}}\left(\frac{\pi_j}{a_j - a_{j-1}}\right) \tag{A62}$$

$$= 1 - \frac{a_J}{\sum_{j=1}^{J} \breve{\pi}_j + \left(1 - \sum_{j=1}^{J} \breve{\pi}_j\right)\sum_{j=1}^{J} D_j + \sum_{j=1}^{J} S_{j,J+1}}$$

$$\times \min_{j\in\{1,\ldots,J+1\}}\left[\frac{\left(1 - \sum_{j'=1}^{J} D_{j'}\right)\breve{\pi}_j + D_j + \sum_{j'>j} S_{j,j'} - \sum_{j'<j} S_{j',j}}{a_j - a_{j-1}}\right] \tag{A63}$$

$$\geq 1 - \frac{a_J}{\sum_{j=1}^J \breve{\pi}_j + \left(1 - \sum_{j=1}^J \breve{\pi}_j\right) \sum_{j=1}^J D_j^* + \sum_{j=1}^J S_{j,J+1}^*}$$

$$\times \min_{j \in \{1,\ldots,J+1\}} \left[\frac{\left(1 - \sum_{j'=1}^J D_{j'}^*\right) \breve{\pi}_j + D_j^* + \sum_{j'>j} S_{j,j'}^* - \sum_{j'<j} S_{j',j}^*}{a_j - a_{j-1}} \right] \qquad \text{(A64)}$$

$$= 1 - \frac{a_J}{\sum_{j=1}^J \pi_j^*} \min_{j \in \{1,\ldots,J+1\}} \left(\frac{\pi_j^*}{a_j - a_{j-1}} \right), \qquad \text{(A65)}$$

where we have modified the notation slightly for this proof, letting uppercase D_j denote deletion from subinterval j and $S_{j,j'}$ denote the shift from j to j', reserving lowercase d_j and $s_{j,j'}$ for the normalized version of those removals, introduced later. Equation (A63) follows from making two substitutions. For the first substitution, we obtain an expression to substitute for π_j appearing in parentheses in (A62) by substituting for $\breve{m}$ from (A49) into (A50) and rearranging:

$$\pi_j = \left(1 - \sum_{j'=1}^J D_{j'}\right) \breve{\pi}_j + D_j + \sum_{j'>j} S_{j,j'} - \sum_{j'<j} S_{j',j}. \qquad \text{(A66)}$$

For the second substitution, we obtain an expresion to substitute for the sum $\sum_{j=1}^J \pi_j$ appearing in the denominator in (A62) by summing (A50) over $j = 1, \ldots, J$ and substituting for $\breve{m}$ from (A49):

$$\sum_{j=1}^J \pi_j = \sum_{j=1}^J \breve{\pi}_j + \left(1 - \sum_{j=1}^J \breve{\pi}_j\right) \sum_{j=1}^J D_j + \sum_{j=1}^J S_{j,J+1}. \qquad \text{(A67)}$$

Inequality (A64) follows from substituting starred variables for their analogues appearing in (A63), where the starred variables are the nonnegative values minimizing (A63). Equation (A65) reflects the shorthand notation

$$\pi_j^* = \left(1 - \sum_{j'=1}^J D_{j'}^*\right) \breve{\pi}_j + D_j^* + \sum_{j'>j} S_{j,j'}^* - \sum_{j'<j} S_{j',j}^*, \qquad \text{(A68)}$$

which is the same proportion as in (A66) after substituting optimal deletions D_j^* and sifts $S_{j',j}^*$.

The derivative of (A64) with respect to D_j^* depends on whether j is a pivotal interval, establishing the minimum in square brackets. If the expression for j is strictly above the minimum, then an increase in D_j^* has no effect on the minimum in square brackets but increases the denominator of the last factor, increasing the overall expression. Hence, $D_j^* = 0$ in that case. If j is pivotal, then the left-hand derivative of (A63) equals

$$\frac{-a_J}{a_j - a_{j-1}} \left[\frac{(1 - \breve{\pi}_j) \sum_{j'=1}^J \pi_{j'}^* - \left(1 - \sum_{j'=1}^J \breve{\pi}_{j'}\right) \pi_j^*}{\left(\sum_{j'=1}^J \pi_{j'}^*\right)^2} \right]. \qquad \text{(A69)}$$

Equation (A69) is negative for $j \leq J$ since the first factor is negative and the numerator and denominator in square brackets are positive. That the numerator in square brackets is positive follows from $\sum_{j'=1}^{J} \pi_{j'}^* \geq \pi_j^*$ and $\sum_{j'=1}^{J} \breve{\pi}_{j'} \geq \breve{\pi}_j$, implying $1-\breve{\pi}_j \geq 1-\sum_{j'=1}^{J} \breve{\pi}_{j'}$. Thus, the optimum for all $j \leq J$ involves increasing D_j until j is no longer pivotal. This leaves interval $J+1$ as pivotal since there is no removal in this interval, so $D_{J+1}=0$.

Reflecting the pivotalness of interval $J+1$, (A64) becomes

$$1-\left(\frac{a_J}{1-\breve{\pi}_{J+1}+\breve{\pi}_{J+1}\sum_{j=1}^{J} D_j^* + \sum_{j=1}^{J} S_{j,J+1}^*}\right)\left(\frac{\breve{\pi}_{J+1}-\breve{\pi}_{J+1}\sum_{j=1}^{J} D_j^* - \sum_{j=1}^{J} S_{j,J+1}^*}{1-a_J}\right) \tag{A70}$$

$$=1-\left(\frac{a_J}{1-\breve{\pi}_{J+1}+\breve{\pi}_{J+1}\sum_{j=1}^{J} D_j^*}\right)\left[\frac{\breve{\pi}_{J+1}\left(1-\sum_{j=1}^{J} D_j^*\right)}{1-a_J}\right], \tag{A71}$$

where (A70) holds because (A65) is increasing in $S_{j,J+1}^*$, implying $S_{j,J+1}^*=0$. Since (A70) is increasing in D_j^*, the minimizing value is either 0 or the positive value that forces a tie between j and $J+1$ for value of the factor in square brackets in (A64), i.e.,

$$\frac{\left(1-\sum_{j'=1}^{J} D_{j'}^*\right)\breve{\pi}_j + D_j^* - \sum_{j'<j} S_{j',j}^* + \sum_{j'>j} S_{j,j'}^*}{a_j - a_{j-1}} = \frac{\left(1-\sum_{j'=1}^{J} D_{j'}^*\right)\breve{\pi}_{J+1}}{1-a_J}. \tag{A72}$$

Solving (A72) for D_j^* yields

$$D_j^* = \left(1-\sum_{j'=1}^{J} D_{j'}^*\right)\left(\delta_j + \sum_{j'=1}^{j-1} s_{j',j}^* - \sum_{j'=j+1}^{J} s_{j,j'}^*\right), \tag{A73}$$

defining

$$\delta_j = \left(\frac{a_j - a_{j-1}}{1-a_J}\right)\breve{\pi}_{J+1} - \breve{\pi}_j \tag{A74}$$

as well as the normalized version of a shift:

$$s_{j,j'}^* = \frac{S_{j,j'}^*}{1-\sum_{j'=1}^{J} D_{j'}^*}. \tag{A75}$$

Combining the two possible values for D_j^* in a single expression yields

$$D_j^* = \left(1-\sum_{j'=1}^{J} D_{j'}^*\right)\max\left(0, \delta_j + \sum_{j'=1}^{j-1} s_{j',j}^* - \sum_{j'=j+1}^{J} s_{j,j'}^*\right), \tag{A76}$$

where the $J(J-1)/2$ variables

$$\{s_{j',j}^* \geq 0 \,|\, j, j' = 1, \ldots, J; j > j'\} \tag{A77}$$

are set to minimize the sum appearing in (A71):

$$\sum_{j=1}^{J} D_j^* = \sum_{j=1}^{J} \max\left[0, \delta_j + \sum_{j'=1}^{j-1} s_{j',j}^* - \sum_{j'=j+1}^{J} s_{j,j'}^*\right], \tag{A78}$$

as it can be verified that minimizing that sum minimizes the overall expression. The variables in (A77) shift mass across intervals in the removal region leaving the total mass constant. Since the max operator in (A78) is convex, the sum is minimized by reducing variation among the D_j^*. If mass could be shifted in both directions, it would be possible to eliminate all variation in the D_j^*. Since the nature of p-hacking only entails shifts in one direction, the best that can be done to achieve the minimum is to eliminate the variation in D_j^* in the form of increases from D_{j-1}^* to $D_j^* > D_{j-1}^*$.

We label the iterative procedure to do so "backward ironing," drawing a connection to a related procedure familiar from the mechanism-design literature (see, e.g., Myerson 1981) for ironing out nonmonotonicities in the distribution of agents' virtual values as a preliminary analytical step. Among other differences, ironing shifts mass in two directions in the mechanism-design literature, whereas the nature of p-hacking constrains shifts to go in one direction here.

To gain intuition for the backward ironing procedure, consider the example in Figure A1 in which the removal region is divided into five subintervals. For pedagogical purposes, we take the subinterval widths to be equal, reducing the problem of equalizing the d_j^* to the problem of equalizing the heights of the shaded bars. Stage 1 shows the density function before any ironing. In stage 2, mass that was shifted by p-hacking from bar 1 to bar 2 is returned to bar 1, equalizing their heights. In stage 3, there is no way to use a backward shift to equalize the height of bar 3 and preceding bars since the preceding bars are higher; so no shift is made. Stage 4 equalizes the heights of bars 3 and 4 by returning mass that was shifted from 3 to 4 back to 3. Stage 5 equalizes the height of bar 5 not only with bar 4, its immediate predecessor, but also with bar 3. Equal mass is returned to both, implying that the drop in the height of bar 4 is twice the rise in bars 3 and 4. More generally, backward ironing redistributes mass to as many predecessors as possible—all those tied with the bar's immediate predecessor after the previous stage's smoothing. Since bar 5 is lower than the stage-2 levels of bars 1 and 2, there is no way to go even further backward and equalize the height of those bars with a backward shift; so backward ironing in stage 5 stops after bar 5 is equalized with bars 3 and 4.

To obtain the formulas in the statement of the proposition requires a change of variables. Let

$$d_j^* \equiv \max\left(0, \delta_j + \sum_{j'=1}^{j-1} s_{j',j}^* - \sum_{j'=j+1}^{J} s_{j,j'}^*\right), \tag{A79}$$

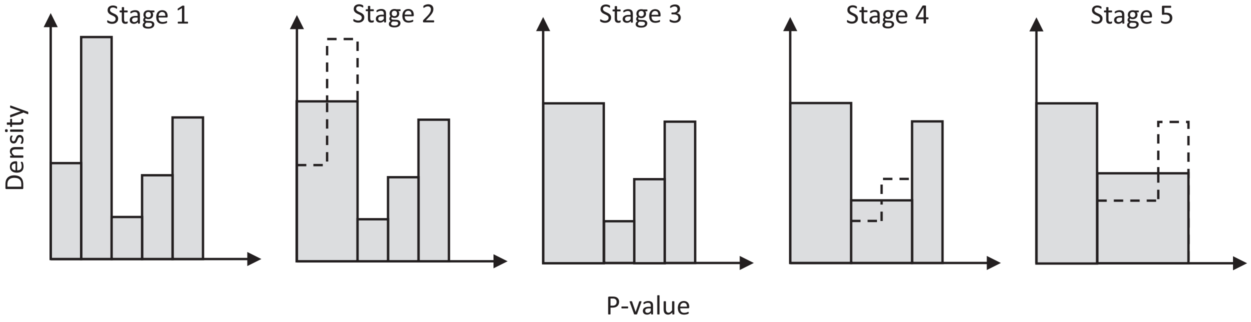
implying

$$\sum_{j=1}^{J} D_j^* = \frac{\sum_{j=1}^{J} d_j^*}{1 + \sum_{j=1}^{J} d_j^*}. \tag{A80}$$

Substituting for the left-hand sum from the right-hand side of (A80) into (A71) and rearranging yields the bound formula (A52) as stated in the proposition. □

Though the bound in Proposition 5 requires a complex series of recursive formulas to state

Figure A1: Example of Backward Ironing



algebraically, the implied algorithm for obtaining the bound is easier to describe. First, the algorithm iteratively shifts mass backward just within the removal region to effectively iron out any divots of mass there. The algorithm then considers a final round of deletions to bring densities in removal-region subintervals up to the density in the no-removal region. Remaining excess mass can then be attributed to misspecified studies and the bound computed by dividing this excess mass by the mass in the removal region adjusted for deletions.

The next proposition provides a possible improvement to the lower bound on removal when moving from a single to multiple removal regions. As with Proposition 5, the bound is complex to formalize algrebraically but easier to describe algorithmically. The algorithm is similar to the one just described but with subtle differences. The algorithm starts out with the same first step of iteratively ironing backward to smooth out any divots of mass within the removal region. However, instead of the final round using deletions to bring densities in removal-region subintervals up to the no-removal region, this is accomplished with shifts. The mass in that final round of shifts bounds the minimal amount of removal in the data.

The proof of the proposition relies on Theorem 1 of Elliot, Kudrin, and Wüthrich (2022), which states that the p-curve is nonincreasing under general conditions, allowing for general misspecification and allowing the p-values to come from general tests (not just t-tests). Their result is stronger than the one established in equation (4), which states that under general misspecification, the p-curve is first order stochastic dominated by the uniform distribution.

Proposition 6. *Suppose we, the researchers, are uninformed about the misspecification rate μ_r. There exists no admissible configuration of removals $\{d_j, s_{j,j'} \,|\, j = 1,\ldots,J; j' > j\}$ such that the removal rate*

$$\rho_r = \frac{\sum_{j=1}^{J}(d_j + s_{j,J+1})}{\sum_{j=1}^{J}\pi_j}, \tag{A81}$$

falls below lower bound

$$\underline{\rho}_r = \frac{\sum_{j=1}^{J} s^*_{j,J+1}}{\sum_{j=1}^{J}\breve{\pi}_j + \sum_{j=1}^{J} s^*_{j,J+1}}, \tag{A82}$$

for $s^*_{j,J+1}$ defined in (A55).

Proof. We begin by translating the monotonicity result in Theorem 1 of Elliot, Kudrin, and Wüthrich (2022) into a constraint on the removal configuration $\{d_j, s_{j,j'} \,|\, j = 1,\ldots,J; j' > j\}$ that can be added to the other admissibility constraints considered so far in the paper. Consider two subintervals, j and $j' > j$. Letting f denote the density of p-values in the population of studies, the monotonicity result implies

$$\begin{aligned}
\frac{\pi_j}{a_j - a_{j-1}} &= \frac{1}{a_j - a_{j-1}} \int_{a_{j-1}}^{a_j} f(p)dp && \text{(A83)}\\
&\geq \frac{1}{a_j - a_{j-1}} \int_{a_{j-1}}^{a_j} f(a_j)dp && \text{(A84)}\\
&= f(a_j) && \text{(A85)}\\
&\geq f(a_{j'}) && \text{(A86)}\\
&= \frac{1}{a_{j'} - a_{j'-1}} \int_{a_{j'-1}}^{a_j} f(a'_j)dp && \text{(A87)}\\
&\geq \frac{1}{a_{j'} - a_{j'-1}} \int_{a_{j'-1}}^{a_j} f(p)dp && \text{(A88)}\\
&\frac{\pi_{j'}}{a_{j'} - a_{j'-1}}. && \text{(A89)}
\end{aligned}$$

In the particular case in which j' is taken to be $J+1$, (A83)–(A89) imply

$$\frac{\pi_j}{a_j - a_{j-1}} \geq \frac{\pi_{J+1}}{1 - a_J}. \tag{A90}$$

A lower bound on (A81) can be obtained by choosing removal configuration $\{d_j, s_{j,j'} \,|\, j = 1, \ldots, J; j' > j\}$ to minimize (A81) subject to admissibility and monotonicity constraints. We proceed to solve a relaxed problem that generates a solution satisfying the original problem, as can be verified. To arrive at the relaxed problem, we let the smaller set of monotonicity constraints embodied in (A90) for $j = 1, \ldots, J$ stand in for the larger set in (A83)–(A89) for all $j' > j$ and $j = 1, \ldots, J$. Further, we extract the information in (A49) and (A50) by substituting for $\breve{m}$ from (A49) into (A50), substituting the resulting value of π_j in the objective function and constraints (A90), and then ignoring constraints (A49) and (A50) in the subsequent minimization. The problem reduces to minimizing the rewritten objective

$$\frac{\sum_{j=1}^J (d_j + s_{j,J+1})}{\left(1 - \sum_{j=1}^J d_j\right) \sum_{j=1}^J \breve{\pi}_j + \sum_{j=1}^J \sum_{j=1}^J (d_j + s_{j,J+1})} \tag{A91}$$

subject to (A47), (A48), and, for all $j = 1, \ldots, J$,

$$\frac{\left(1 - \sum_{j'=1}^J d_{j'}\right) \breve{\pi}_j + d_j + \sum_{j'>j} s_{j,j'} - \sum_{j'<j} s_{j',j}}{a_j - a_{j-1}} \geq \frac{\left(1 - \sum_{j'=1}^J d_{j'}\right) \breve{\pi}_{J+1} - \sum_{j=1}^J s_{j,J+1}}{1 - a_J}. \tag{A92}$$

We will show that the minimum involves all shifts, no deletions. To this end, consider a removal configuration in which $d_j > 0$. We will show this configuration can be improved upon by one that reduces d_j by Δd_j and increases $s_{j,J+1}$ by Δs_j such that the changes satisfy

$$\Delta d_j \left(\frac{\breve{\pi}_{J+1}}{1-a_J} - \frac{1-\breve{\pi}_j}{a_j - a_{j-1}} \right) = \Delta s_j \left(\frac{1}{a_j - a_{j-1}} - \frac{1}{1-a_J} \right). \tag{A93}$$

By construction, Δd_j and Δs_j keep the gap between the left- and right-hand sides of constraint (A92) the same, so the constraint continues to be satisfied. The sum

$$\sum_{j=1}^{J} (d_j + s_{j,J+1}) \tag{A94}$$

changes by $\Delta s_j - \Delta d_j$, which upon substituting for Δs_j from (A93) and rearranging becomes

$$-\frac{(a_j - a_{j-1})(1-\breve{\pi}_{J+1}) + (1-a_J)\breve{\pi}_j}{(1-a_J) + (a_j - a_{j-1})}, \tag{A95}$$

which is obviously negative. The new configuration not only reduces the sum (A94); we can also prove that it reduces the objective (A91). Objective (A91) is increasing in (A94), which appears as a term in the numerator and denominator of (A91). Thus, the reduction in (A94) reduces (A91). The reduction in d_j reduces (A91) through an additional channel: d_j appears outside of the sum (A94) in the factor $(1 - \sum_{j=1}^{J} d_j) \sum_{j=1}^{J} \breve{\pi}_j$ in the denominator; a reduction in d_j reduces (A91) through this channel. Having demonstrated that the new configuration reduces the objective, the original configuration with $d_j > 0$ could not have been optimal, implying that the minimum involves $d_j^* = 0$.

Substituting $d_j^* = 0$ into (A91), the objective function becomes

$$\frac{\sum_{j=1}^{J} s_{j,J+1}}{\sum_{j=1}^{J} \breve{\pi}_j + \sum_{j=1}^{J} \sum_{j=1}^{J} s_{j,J+1}}. \tag{A96}$$

Minimizing (A96) is equivalent to minimizing the sum

$$\sum_{j=1}^{J} s_{j,J+1}. \tag{A97}$$

Constraint (A92) becomes

$$\frac{\breve{\pi}_j + \sum_{j'>j} s_{j,j'} - \sum_{j'<j} s_{j',j}}{a_j - a_{j-1}} \geq \frac{\breve{\pi}_{J+1} - \sum_{j=1}^{J} s_{j,J+1}}{1-a_J}. \tag{A98}$$

If constraint (A98) is slack for j, then $s_{j,J+1}^* = 0$. If (A98) binds for subinterval j and is slack for subinterval $j' > j$, the solution can be improved by increasing $s_{j,j'}$ and relaxing the constraint for j. Indeed, $s_{j,j'}$ should be increased up until the point that (A98) binds for j'. This proves that if constraint (A98) binds for any j, it binds for all $j' = j, \ldots, J$.

The minimizing solution can be found by backward ironing among subintervals in the removal region until completion and then one final step of backward ironing from the no-removal region back to the removal region. The formulas for the shifts involved in backward ironing are provided in Proposition 5. □

Estimating Improved Bounds on Misspecification

This section estimates the lower bound on misspecification, $\underline{\mu}_r$ from equation (A51), in the pure subsample of RCT balance tests and pure subsample of other tests in our data. Instead of using the estimate $\hat{\pi}_r$ of the proportion of p-values aggregated over the removal region $[0, 0.15)$ from one of the lower rows in Table 1, we will use estimates $\hat{\pi}_1$, $\hat{\pi}_2$, and $\hat{\pi}_3$ for the finer subintervals from the top rows in the table, where $a_1 = 0.05$, $a_2 = 0.10$, and $a_3 = 0.15$, implying that $\hat{\pi}_1$, $\hat{\pi}_2$, and $\hat{\pi}_3$ are the estimated proportion of observations in the subintervals $[0, 0.05)$, $[0.05, 0.10)$, and $[0.10, 0.15)$, respectively.

Start by computing $\underline{\mu}_r$ for the pure sample of RCT balance tests. To provide geometric intuition for the computations, Figure A2 graphs the regression results as shaded bars. The top panel provides the relevant results for the pure sample of RCT balance tests. The first step is to apply backward ironing. The only scope for backward ironing in the removal region is to equalize the heights of bars 2 and 3. We have $d_1^* = \delta_1$, $d_2^* = \delta_2^* + s_{2,3}^*$, and $d_3^* = \delta_3 - s_{3,2}^*$, implying $d_1^* + d_2^* + d_3^* = \delta_1 + \delta_2 + \delta_3$, in turn implying $\sum_{j=1}^J d_j^* = \sum_{j=1}^J \delta_j$ in this case. Hence,

$$\sum_{j=1}^{J} d_j^* = \sum_{j=1}^{J} \delta_j = \left(\frac{a_J}{1-a_J}\right) \breve{\pi}_{J+1} - \sum_{j=1}^{J} \breve{\pi}_j = \frac{1}{1-a_J}\left(a_j - \sum_{j=1}^{J} \breve{\pi}_j\right), \tag{A99}$$

where the first equality was just established, the second follows from (A74), and the last from algebra. Substituting from (A99) for $\sum_{j=1}^J d_j^*$ in (A51) leads to a 0 in the numerator, implying $\underline{\mu}_r = 0$.

This is the same (trivial) bound $\underline{\mu}_r$ found before with a single removal region. Moving to multiple removal regions does not tighten the misspecification bound for the pure sample of RCT balance tests. We still cannot rule out no misspecification. The p-curve in the top panel of Figure A2 could have corresponded to the uniform distribution in the pre-removal sample from which δ_1, δ_2, and δ_3 was deleted.

Turn to computing $\underline{\mu}_r$ for the pure sample of other tests. The shaded bars in the lower panel of Figure A2 represent the regression results for this sample. There is no scope for backward ironing in the removal region in this case, so we can proceed directly to computing the deletions $d_1^* = \max(0, \delta_1) = 0$, $d_2^* = \max(0, \delta_2) = \delta_2$, and $d_3^* = \max(0, \delta_3) = \delta_3$. Hence,

$$\sum_{j=1}^{J} d_j^* = \delta_2 + \delta_3 = \frac{0.10}{0.85}(1 - \breve{\pi}_1 - \breve{\pi}_2 - \breve{\pi}_3) - \breve{\pi}_2 - \breve{\pi}_3, \tag{A100}$$

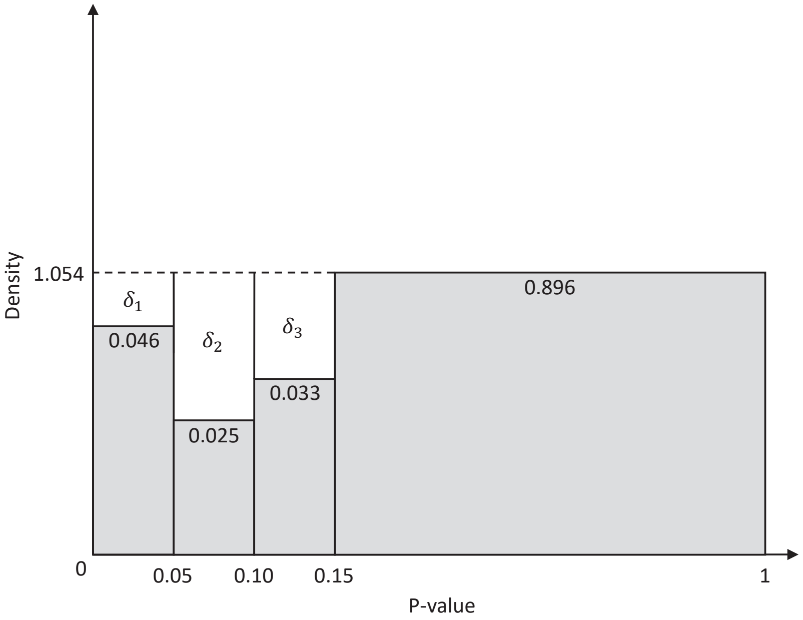
where the second equality follows from (A74), noting we are taking the threshold for the removal region to be $a_J = 0.15$. Substituting from (A100) for $\sum_{j=1}^J d_j^*$ in (A51) and rearranging yields

$$\underline{\mu}_r = \frac{0.90\breve{\pi}_1 - 0.05(1 - \breve{\pi}_2 - \breve{\pi}_3)}{0.75\breve{\pi}_1 + 0.10(1 - \breve{\pi}_2 - \breve{\pi}_3)}. \tag{A101}$$

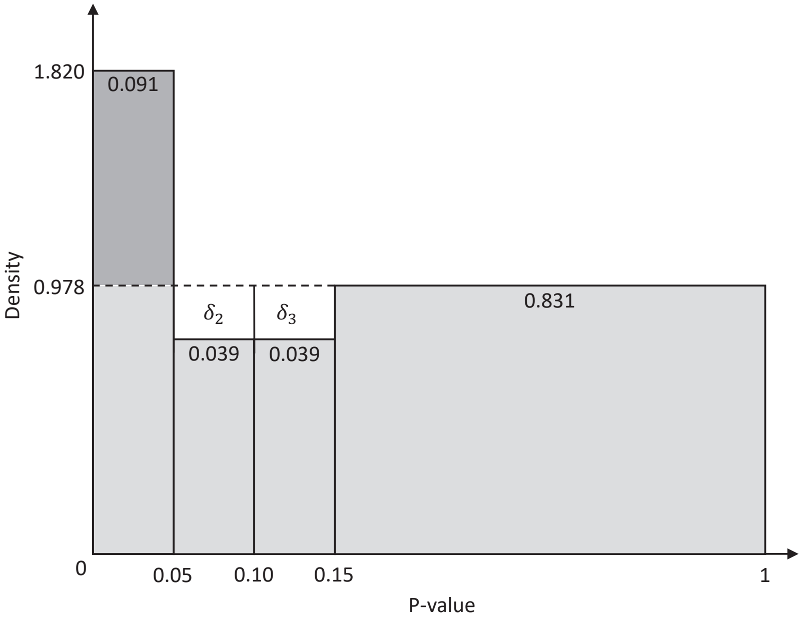
Substituting the estimates of $\hat{\pi}_1$, $\hat{\pi}_2$, and $\hat{\pi}_3$ provided in the Table 1 and the bottom panel of

Figure A2: Computing Improved Bounds on Misspecification with Multiple Removal Regions

(1) RCT balance tests, pure



(2) Other tests, pure



Notes: Graphs of regression estimates for proportions of p-values in various intervals from Table 1. The top panel graphs results from column (2) of the table and the bottom panel from column (3). The area of each shaded bar (including, where relevant, light and dark shaded pieces) corresponds to the proportion of observations in the relevant interval, also indicated by the number at the top of the bar. The non-removal region $[0.15, 1]$ has been consolidated into a single bar. Horizontal axis not drawn to scale: the distance between 0.15 and 1 has been shortened to preserve legibility.

Figure A2, we obtain estimate of the lower bound on the misspecification rate of 0.224. The robust standard error on this estimate, clustered on article, computed using the delta method, is 0.029. This estimate is about 9 percentage points tighter than the bound estimated treating the removal region as a single interval reported in Table 3: an estimate of 0.136 (standard error 0.052). Both estimates are looser than the bound assuming the removal rate for other tests is the same as for RCT balance tests reported in Table 3: an estimate of 0.402 (standard error 0.075).

Estimating Improved Bounds on Removal

This section estimates the lower bound on removal, $\underline{\rho}_r$ from equation (A82), in the pure subsample of RCT balance tests and pure subsample of other tests in our data. For the pure subsample of RCT balance tests, one can show that the backward ironing with multiple removal regions does not tighten the bound relative to the 0.308 estimated in Table 2 for a single removal region.

For Proposition 6 to tighten the bound on removal relative to that with a single removal region, one can show that the p-value distribution has to look like the bottom panel of Figure A2 rather than the top panel. That is, the distribution must have subintervals with shaded bars both above and below that of the no-removal region rather than all the shaded bars in the removal region being below that in the no-removal region as in the top panel. The figure suggests that there is scope for Proposition 6 in the pure sample of other tests, and indeed there is.

The algorithm implied by Proposition 6 involves backward ironing within the removal region and then a round of backward ironing from no-removal to removal region. There is no scope for backward ironing within the removal region because the bars in the removal region are no higher than their counterparts to the left. So we are left to iron backward from the no-removal region to equalize the height of that bar with the those in the intervals $[0.05, 0.10)$ and $[0.10, 0.15)$. Backward ironing entails shifts solving

$$\frac{\breve{\pi}_2 + s^*_{2,4}}{a_2 - a_1} = \frac{\breve{\pi}_3 + s^*_{3,4}}{a_3 - a_2} = \frac{\breve{\pi}_4 - s^*_{2,4} - s^*_{3,4}}{1 - a_J}. \tag{A102}$$

Substituting $a_2 - a_1 = a_3 - a_2 = 0.05$ and $1 - a_J = 0.85$ and solving yields

$$s^*_{2,4} = \frac{0.05}{0.95}(1 - \breve{\pi}_1) - \breve{\pi}_2 \tag{A103}$$

$$s^*_{3,4} = \frac{0.05}{0.95}(1 - \breve{\pi}_1) - \breve{\pi}_3. \tag{A104}$$

Substituting these values of $s^*_{2,4}$ and $s^*_{3,4}$ along with $s^*_{1,4} = 0$ into (A82) yields

$$\underline{\rho}_r = \frac{0.10(1 - \breve{\pi}_1) - 0.95(\breve{\pi}_2 + \breve{\pi}_3)}{0.10 + 0.85\breve{\pi}_1}. \tag{A105}$$

Substituting the estimates of $\hat{\pi}_1$, $\hat{\pi}_2$, and $\hat{\pi}_3$ provided in the Table 1, we obtain estimate of the lower bound on the removal rate of 0.092 with a robust standard error, clustered on article, of 0.028. This estimate is significantly greater than 0 in a one-tailed test at better than the 1% level but does not amount to a substantial improvement over the trivial 0 bound with a single removal region.

Additional References

Myerson, Roger. B., “Optimal Auction Design,” *Mathematics of Operations Research* 6:1 (1981), 58–73.

Pounds, Stan and Stephan W. Morris, “Estimating the Occurrence of False Positives and False Negatives in Microarray Studies by Approximating and Partitioning the Empirical Distribution of P-Values,” *Bioinformatics* 19:10 (2003), 1236–1242.

A5. Inverse Frequency Weighting

This appendix provides additional discussion of the biases addressed by inverse frequency weighting used in our regression specifications.

As explained in the main text, we focus on conditional expectation (21), corresponding to the mean outcome of a randomly selected sniff test from a randomly selected table. An obvious alternative would be to pool sniff and look at the mean outcome for a random selection from that pool, captured by the expected value

$$E_{i\in I_j}(\mathbf{1}(p_i \in j)). \tag{A106}$$

If different tables articles contain different numbers of sniff tests and the number of sniff tests in the table is correlated with their significance, then expectation (A106) will differ from (21). Pooling sniff tests together and taking a simple unweighted mean will lead tables reporting more sniff tests to be over-represented.

We will prove that the unweighted mean is a biased estimate of (21) in a predictable direction. Let $\bar{n}_j = E_{t\in T_j}(|t|)$ denote the expected number of observations per table in the set T_j. By the law of iterated expectations, the unweighted expectation (A106) can be written

$$E_{t\in T_j}\left(\frac{|t|}{\bar{n}_j}E(\mathbf{1}(p_i \in j)\,|\, i\in t)\right). \tag{A107}$$

It is immediate that if tables are all the same size, then (21) equals (A107) since $|t| = \bar{n}_j$.

To provide a more general comparison of the two expectations, we introduce the following reduced-form model. Repeating equation (25), we have

$$\mathbf{1}(p_i \in j) = \breve{\pi}_j + u_{ij}, \tag{A108}$$

where u_{ij} is an error term with conditional expectation

$$E_{t\in T_j}(E(u_{ij}\,|\, i\in t)) = 0 \tag{A109}$$

by (21). Assume the error term can be decomposed as $u_{ij} = \nu_{ij} + \varepsilon_{ij}$, where ε_{ij} is white noise with mean zero both within a table and across all tables, i.e.,

$$E(\varepsilon_{ij}\,|\, i\in t) = 0, \tag{A110}$$

whereas ν_{ij} is unobserved table effect, which has zero mean across tables but can have nonzero mean within a table. Formally, letting η_{tj} denote expected value of the unobservable effect for table t, i.e.,

$$\eta_{tj} = E(\nu_{ij} \,|\, i \in t), \tag{A111}$$

we allow η_{tj} to be nonzero.

Substituting (A108)–(A111) into (A107) and rearranging yields

$$\frac{1}{\bar{n}_j} E_{t \in T_j} \left(E(|t|\breve{\pi}_j + |t|\nu_{ij} + |t|\varepsilon_{ij} \,|\, i \in t) \right) = \breve{\pi}_j + \frac{1}{\bar{n}_j} \mathrm{Cov}_{t \in T_j}(|t|\eta_{tj}). \tag{A112}$$

We see that the alternative expectation is biased relative to π_j: in the positive direction if η_{tj} covaries positively with table size $|t|$ and in the negative direction if η_{tj} covaries negatively with $|t|$.

The main text showed that inverse frequency weighting produces an unbiased estimator of (21). Wooldridge (2010, chapter 20) discusses the utility of weighting across a variety of sampling contexts. Connections can be drawn between our weighted estimators those prescribed in Wooldridge's text. Our equation (22) implements the estimator suggested in his equation (20.48) for for cluster-sampling contexts. Equation (24) implements the estimator suggested in his equation (20.13) for standard-stratified-sampling context. 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.

A6. Supplementary Regression Results

Table A8: Regression Results for Proportion of Significant P-values

		RCT balance tests		Other tests		
	Full sample	Pure	Possibly improved	Pure	Matched, pre-match	Matched, post-match
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Proportion $\hat{\pi}_j$ of p-values in subintervals						
• $j = [0, 0.05)$	0.099^{***}	0.046	0.057	0.091^{***}	0.560^{***}	0.041
	(0.006)	(0.009)	(0.011)	(0.006)	(0.039)	(0.011)
• $j = [0.05, 0.10)$	0.038^{***}	0.025^{***}	0.037^{***}	0.039^{***}	0.070^{*}	0.024^{***}
	(0.002)	(0.005)	(0.005)	(0.003)	(0.010)	(0.006)
• $j = [0.10, 0.15)$	0.037^{***}	0.033^{***}	0.039^{**}	0.039^{***}	0.029^{***}	0.031^{**}
	(0.003)	(0.006)	(0.005)	(0.004)	(0.007)	(0.009)
• $j = [0.15, 0.20)$	0.047	0.054	0.038^{***}	0.050	0.034^{**}	0.044
	(0.004)	(0.009)	(0.004)	(0.007)	(0.007)	(0.008)
Aggregate proportion $\hat{\pi}_r$	0.173^{***}	0.103^{***}	0.132	0.170^{**}	0.660^{***}	0.096^{***}
over removal region $[0, 0.15)$	(0.007)	(0.012)	(0.005)	(0.009)	(0.034)	(0.018)
Observation counts						
• Unique sniff tests	28,832	2,953	4,020	16,063	2,372	3,424
• Clusters	857	85	109	606	71	109

Notes: Rounding out the partial set of regression results provided by Table 1, this table provides the full set of results corresponding to all panels of Figure 3. Notes from Table 1 apply.

Table A9: Strength of Author Claims

	RCT balance tests, pure			Other tests, pure		
	$p_i \in [0, 0.15)$	$p_i \in [0.15, 1]$	Difference	$p_i \in [0, 0.15)$	$p_i \in [0.15, 1]$	Difference
Variable	(1)	(2)	(1) – (2)	(3)	(4)	(3) – (4)
Nature of claim						
• Strong	0.775***	0.939***	−0.164***	0.517***	0.852***	−0.335***
	(0.082)	(0.027)	(0.060)	(0.055)	(0.024)	(0.048)
• Weak	0.065	0.029	0.036	0.096***	0.051***	0.045**
	(0.045)	(0.020)	(0.025)	(0.029)	(0.017)	(0.021)
• Admit rejected	0.160**	0.032*	0.128**	0.309***	0.058***	0.251***
	(0.076)	(0.019)	(0.060)	(0.054)	(0.012)	(0.049)
• No claim	†	†	†	0.078**	0.039***	0.039
				(0.026)	(0.014)	(0.027)
Observation counts						
• Sniff tests	261	1,246	1,507	1,056	3,606	4,662
• Clusters	39	58	58	139	226	226

Notes: Results are the proportions of p-value observations out of the subsample in the interval and sample indicated in column headings. The numbers of observations are different from those reported in Figure 2 because we exclude here the observations for which we cannot determine whether the p-value falls in $[0, 0.15)$. The type of the author claim is indicated in the row heading. Columns (1) and (3) focus on p-values in the removal region $[0, 0.15)$ and columns (2) and (4) on p-values above the removal region. Since they are proportions of the indicated subsample, results in each of columns (1)–(4) add down the column to 1. Results are from IFWLS regressions. Standard errors, reported in parentheses below results, are clustered at the article level. †No observations in the pure sample of RCT balance tests categorized as making no claim. Significantly different from 0 at the *ten-percent level, **five-percent level, ***one-percent level. The relevant test is one-tailed in columns (1)–(4) and two-tailed in the remaining columns, which report differences.