NBER WORKING PAPER SERIES

SNIFF TESTS AS A SCREEN IN THE PUBLICATION PROCESS: THROWING OUT THE WHEAT WITH THE CHAFF

Christopher Snyder
Ran Zhuo

Working Paper 25058
http://www.nber.org/papers/w25058

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2018

The authors are grateful to the Sloan Foundation for financial support as well as to three programs at Dartmouth College (Dartmouth Economic Research Scholars, Lucas Family Fund for Undergraduate Research, and Presidential Scholars Program) for supplementary funding for some research-assistant time. The authors thank Barbara DeFelice, Ethan Lewis, Bruce Sacerdote, and Douglas Staiger for advice and suggestions. The authors are indebted to their dedicated team of research assistants: Kelsey Anspach, Ekshesh Bekele, Albert Chen, Barry Chen, Natalia Drozdoff, Isabel Hurley, Paul Jeon, Amrita Misha, Kirill Savolainen, and Olivia Zhao.The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Sniff Tests as a Screen in the Publication Process: Throwing out the Wheat with the Chaff
Christopher Snyder and Ran Zhuo
NBER Working Paper No. 25058
September 2018, Revised July 2020
JEL No. A14,B41,C18

ABSTRACT

The increasing demand for empirical rigor has led to the growing use of auxiliary tests (balance, specification, over-identification, placebo, etc.) in assessing 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. Using a sample of nearly 30,000 published sniff tests collected from scores of economics journals, we study the use of sniff tests as a screen in the publication process. For the subsample of balance tests in randomized controlled trials, our structural estimates suggest that the publication process removes 46% of significant sniff tests, yet only one in ten of these is actually misspecified. For other tests, we estimate more latent misspecifiation and less removal. Surprisingly, more authors would be justified in attributing significant sniff tests to random bad luck.

Christopher Snyder
Department of Economics
Dartmouth College
301 Rockefeller Hall
Hanover, NH 03755
and NBER
chris.snyder@dartmouth.edu

Ran Zhuo
Department of Economics
Harvard University
Cambridge, MA 02138
USA
rzhuo@g.harvard.edu

1. Introduction

The increasing demand for rigor in empirical economics has led to the increasing use of auxiliary tests (balance, specification, over-identification, falsification, placebo, etc.) to assess the credibility of a paper's main results. We dub these tests "sniff tests" because rejection is bad news for the author and standards for passing are informal.

Sniff tests undeniably provide valuable information that, when combined with the paper's main test results, allows readers to better update their priors about the researched subject. Sniff tests may not be completely benign, however, if the publication process uses them as a screen. It is natural to want to discard misspecified studies that would pollute the literature and waste scarce journal space. By chance, 5% of well-specified studies will have probability values (p-values) significant at the 5% level, 10% of them at the 10% level, and so forth, making them appear to be misspecified when judged by those respective thresholds. If the publication process screens out these studies, valuable research is lost. Articles having bad luck with their sniff tests, presumably otherwise of average quality, end up in what Rosenthal (1979) termed the "file drawer," replaced by articles of marginal quality having better luck with their sniff tests. This cost may be worth bearing if enough misspecified studies are also caught by the screen. To understand whether the publication process is operating efficiently, it is useful to measure how many papers are screened out on the basis of sniff tests and the proportions that are well-specified versus misspecified. How much wheat is being thrown out with the chaff?

It would be difficult to answer this question by analyzing a small sample of published papers. We contribute to the literature by painstakingly collecting a sample of 29,755 sniff tests from 892 articles published in 59 economics journals. Under the null hypothesis of no misspecification and absent screening 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 from our large sample of sniff tests to the uniform benchmark allows us to uncover the extent of screening on significant p-values and the extent of misspecification in the underlying set of papers.

We start with a reduced-form analysis, amounting to visual inspection of kernel densities of p-values from various subsamples. While we take an initial look at the full dataset, most of the analysis focuses on what we call the "pure" sample of sniff tests, where we are sure the authors

did not take measures (re-randomization, stratification, or matching) to lessen the p-values' significance. We further break the pure sample down into balance tests in randomized controlled trials (RCTs) and other tests. We expect that few RCTs suffer from flawed randomization procedures, ruling out a substantial proportion of misspecified studies. Unless significant p-values have been removed at some point in the publication process, the aggregate distribution of p-values for balance tests in RCTs should closely resemble the uniform distribution. For other tests, an unknown, perhaps substantial, proportion of studies may suffer from misspecification, skewing the aggregate distribution of p-values toward 0 in the absence of removal.

For the pure sample of balance tests in RCTs, the kernel density of p-values looks relatively uniform, except for a block of missing mass in the $[0, 0.15)$ interval (confirmed by formal tests to be a statistically significant departure from uniformity). This shape is consistent with the balance tests in RCTs not suffering from much misspecification but being subject to a substantial rate of screening for significant p-values. For the pure sample of non-balance tests, the kernel density of p-values 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, enough that considerable mass of p-values is concentrated in the significant range even after some is removed by the publication process.

Authors would like to claim that significant sniff tests are the result of bad luck rather than misspecification, but such claims are easier to evaluate in the aggregate. Using information collected on the narrative authors used to characterize their sniff tests collected from the articles' text, we show that, surprisingly, fewer authors attribute their failed sniff tests to random bad luck than would be justified in doing so.

Moving from reduced-form to structural methods, we specify a two-stage model in which the initial population is a mixture of well-specified studies, whose p-values are characterized by the uniform distribution, and misspecified studies, whose p-values follow an alternative density approximated by a flexible beta distribution. In the second stage, the publication process removes a proportion of studies with significant sniff tests. The model predicts a particular nonmonotonic shape for the distribution of p-values, which turns out to fit the data quite well, lending credibility to the structural estimates. For the pure sample of balance tests in RCTs, our structural estimates imply that 46% of p-values below 0.15 are removed by the publication process, yet only 11% of

p-values in this range are the result of misspecification (failed randomization). The remaining 89% of p-values in this range are from well-specified studies that could have been resuscitated by conditioning the main results on any unbalanced covariates rather than being screened out. For the pure sample of other tests, 25% of p-values below 0.15 are removed by the publication process, while 36% of p-values in this range are the result of misspecification.

Lacking information on the costs of publishing misspecified studies or removing well-specified studies, we are not in a position to conduct a welfare analysis. We estimate that misspecified studies account for fewer than 1% of balance tests in RCTs with p-values above 0.15 and 2% of other tests with p-values above 0.15. Hence, using a 0.15 threshold for removal appears to be sufficient to keep the published literature relatively free of misspecified studies. Given the contrasting outcomes between balance and other tests, if the screening process is efficient for RCTs, it may be too liberal for other studies. On the other hand, if the screening process is efficient for other studies, it may be too severe for RCTs.

2. Literature Review

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 confirm in a Monte Carlo exercise that the aggregate distribution of p-values from balance tests in RCTs is uniform on $[0, 1]$ even if tables report many tests and their outcomes are correlated. The authors analyze a sample of balance tests from articles in development economics. While their study examined 13 articles, our sample includes nearly 900 articles across all fields of economics, allowing us to obtain a broader view of the distribution of p-values in the economics literature and to run formal statistical tests.

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 and Jaravel (2017) propose new tests for pre-trends in event studies. Andrews, Gentzkow, and Shapiro

(2018) propose a measure relating the performance of sniff tests to the 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. Insteading of the null of no misspecification, Bilinski and Hatfield (2020) suggest flipping the perspective and testing against the null that misspecification exceeds some threshold. Roth (2020) shows that screening studies based on violation of parallel trends can exacerbate publication bias in the main results of interest.

The sniff tests analyzed in our paper differ from tests of main effects on which the broader literature on publication bias has focused. Despite this difference, our paper makes two contributions to this broader literature. First, though we highlight the loss of informative research as the main cost of screening out well-specified studies that happen to have significant sniff-test results, Roth (2020) points out that such screening can indirectly exacerbate the bias in test of main effects. Whether this indirect form of publication bias should be a major concern depends on the proportion of well-specified studies that are screened out relative to misspecified studies. Our estimates suggest that the majority of screened studies are well-specified. Second, while the selection pressure exerted by the publication process is the opposite for sniff test as for main tests—selecting for high rather than low p-values—our results can shed light on the potential strength of selection pressures, regardless of direction. We find strong pressure in some subsamples, removing nearly half of tests from the journals for the pure sample of balance tests in RCTs.

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).

Regarding the methods used in our structural estimation, Iyengar and Greenhouse (1988) and Hedges (1992) pioneered the approach of specifying a parametric model of publication removal, estimated via maximum likelihood. Andrews and Kasy (2019) generalized the approach and applied it to replication studies and meta-analyses in economics and psychology. Since we study sniff tests, not tests of main effects, estimation is simpler because we can work with a known uniform distribution under the null of no misspecification, and this null is approximately true for a subsample of our data (pure sample of balance tests in RCTs). We adapted the beta-uniform mixture model, which we use in our structural model to decompose the aggregate distribution of p-values into populations of nulls and alternatives, from the bioinformatics literature (see, e.g., Pounds and Morris 2003, Datta and Datta 2005, and Do *et al*. 2005). Allison *et. al*. (2006) discuss the use of the model to analyze tests from DNA microchips of the differential expression of a large number of genes simultaneously. Estimating a beta-uniform mixture model on the pooled p-values from those tests allows researchers to determine the relative proportion of related versus unrelated genes.

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 ScienceDirect, Elsevier's online database of journal articles. We collected PDF files for all economics articles that were turned up by a search of related keywords such as "balance test," "baseline comparison," "falsification test," "placebo test," "randomization," "validation check," and so on. We supplemented the Elsevier journals with five top-tier, general-interest journals in economics archived on JSTOR (*American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Review of Economic Studies*, *Quarterly Journal of Economics*), performing the same keyword search as on ScienceDirect. As the keywords were relatively uncommon before 2005, we restricted our pool to articles

published from 2005 to 2015.[1]

The research assistants browsed each article in this initial pool, determining whether it contained a table reporting sniff tests. If so, the research assistant collected data on test statistics, p-values, and significance levels reported in the table or tables containing the sniff tests along with relevant table, article, and journal information. Only two-sided tests were included.[2] All work was re-checked by supervising research assistants. We dropped 764 observations that either were not structured as well-defined hypothesis tests with associated p-values or did not provide sufficient information to glean an exact p-value or an interval for it. Dropping two further observations with a p-value exceeding 1 leaves a final dataset of 29,755 sniff-test observations. These observations are reported in 1,367 different tables from 892 articles published in 59 journals. Appendix Table A1 lists the journals and the contribution each makes to the sample. The *Journal of Health Economics* contributes the most observations (16% of the sample), followed by the *Journal of Public Economics* (11%), the *Journal of Development Economics* (10%), *Labour Economics* (8%), and the *American Economic Review* (8%).

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

Figure 2 presents the subsample breakdown by methodology for the dataset. It is important to distinguish among methodologies because different types of tests can have very different dis-

[1] *Quarterly Journal of Economics* 2013-2015 and *Review of Economic Studies* 2015 are unavailable through JSTOR. We performed searches of the same keywords on Google Scholar to obtain the initial pool of articles for these journal-years.

[2] We omitted one-sided tests given the difficulties of analyzing publication bias in the context of one-sided tests, outlined by DeLong and Lang (1992). However, we did notice some peculiar cases in our data collection in which authors used an inconsistent reporting convention, gauging significance levels for their sniff test according to the tight thresholds of a two-sided test but then discussing the test as if it were one-sided. For example, the authors might test for a pre-trend, discover its p-value is significant at the 5% level according to a two-sided test, but then argue that the trend is in the opposite direction from their result of interest (say a downward trend working against a positive jump at a regression discontinuity). If in fact the test cannot be rejected when a negative trend is found, it was a mistake to have used two-sided rather than one-sided significance thresholds. Of course, the one-sided null would not be rejected in this particular test regardless of significance level used because the trend was on the "accept" side of the inequality. The problem raised by this selective use of one- and two-sided tests is in the interpretation of other sniff tests in the article: one-sided tests may have been appropriate for these, too, but rejections glossed over by the use of tighter two-sided thresholds. It is unclear how best to treat these peculiar cases of one-sided tests posing as two-sided. We checked the robustness of all of our results including and excluding them. The results are very similar in magnitude and significance. All results presented in this paper include these observations.

tributions of p-values. Balance tests from randomized controlled trials (RCTs) will have a special place in our analysis because we expect study flaws due to randomization failure are rare; so their p-values should have little deviation from the uniform $[0, 1]$ distribution. We do not have good *ex ante* knowledge of the amount of study flaws in other tests (balance tests from non-RCTs, placebo tests, and various types of falsification and specification tests). The shape of the mixed distribution will depend on the nature of the underlying data and model appearing in the articles in our sample. For example, specification tests will exhibit a concentration of low p-values if many papers in our dataset happen to suffer from misspecification.

Valid techniques exist to improve balance in RCTs relative to what we will label the "pure" case of a single randomization. As discussed in Athey and Imbens (2017), these techniques including re-randomization (randomizing treatment/control selection multiple times until a desired balance in covariates between groups is achieved), stratification (randomizing treatment/control selection within covariate strata), and matching (finding pairs of observations with the same covariates, making one a treatment and the other a control). Tautologically, application of balance-improvement techniques raises the p-values of the subsequent balance tests. If we included balance tests that have been run after balance improvement in our sample of RCT balance tests, the distribution of p-values would no longer be expected to be uniform in the absence of study flaws. Balance-improvement techniques shift some of the mass of low p-values to higher values, creating an upward bias in estimated amount of removal during the publication process. To avoid this problem, we construct a pure subsample for which we can rule out balance improvement, separating items for which balance improvement cannot be ruled out into a different subsample.

Bruhn and McKenzie's (2009) survey of leading researchers who conduct RCTs in developing countries reveals a "mixed bag" of approaches toward balance. While it is not uncommon for researchers to resort to a single randomization (80% did so for at least one past experiment; 40% did so for their most recent experiment), balance-improvement techniques are also widely used. Their review of selected publications suggests that when authors used re-randomization, the process was not described in detail. To account for opaque or possibly missing mentions of re-randomization, we are conservative in our formation of the pure subsample and only include items for which re-randomization can be ruled out 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. Our pure subsample contains 42% of RCT balance test items. For the remaining 58% of RCT balance tests, either re-randomization is not mentioned in the article but cannot definitively be ruled out because authors had access to the baseline data to re-randomize before treatment assignment[3] or balance was definitively improved by application of re-randomization, stratification, and/or matching.[4]

Turning to the subsample of tests other than balance in RCTs, re-randomization and stratification are irrelevant in this context since authors cannot control the experimental procedure. Matching is the remaining, feasible method for p-value improvement, which authors employ by restricting analysis to a matched subsample of their full sample. In 74% of these other tests, no matching, and thus no p-value improvement method, was employed. The rest involve some matching. For 11% of other tests, matching is employed and results of sniff tests prior to matching are reported. For 16% of other tests, matching is likewise employed but results of sniff tests after matching are reported.

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

We also directed the research assistants to collect qualitative information on how authors characterized the outcome of their sniff tests. We defer discussion and analysis of this variable to Section 5.5.

[3]There is less concern about the opacity of the reporting of stratification and matching; perhaps owing to the deliberation involved in their application, they are thought to be reported whenever used. In sum, therefore, observations in our possibly pure sample are definitively not stratified or matched.

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

4. Reduced-Form Analysis

Our reduced form methods offer a simple descriptive analysis of the amount of deviation of aggregate distributions of p-values away from the uniform benchmark. Section 4.1 presents kernel density estimates for various subsamples. Section 4.2 describes the regression-based specifications used. Section 4.3 presents the regression results and simple statistical tests of deviation from the uniform distribution benchmark.

4.1. Kernel Density Plots

In this subsection, rather than focusing on the mass of p-values in a particular interval, we plot kernel density estimates of the entire distribution. While kernel density estimation is typically applied to random variables with unbounded supports, the support for p-values is $[0, 1]$, having both lower and upper bounds. The kernel density estimator must be modified to account for these bounds.[5] The kernel densities are estimated using the subsample of exact p-values rather than intervals. To account for the fact that different articles contribute different number of sniff tests, we weight each p-value by the inverse of the number of observations in the containing article so each article contributes equally to the aggregate distribution. A more detailed discussion of this weighting scheme can be found in the next subsection where we present technical details behind the regression specifications.

The kernel density plot in Panel (1) of Figure 3 shows all exact p-values in our full sample. One immediately sees the distribution deviates considerably from the uniform, suggesting a sizable number of published sniff tests pick up study flaws. The shape of the kernel density also follows a peculiar nonlinear shape. The plot spikes for the lowest p-values. The mountain of extra mass for p-values below 0.05 is taken from the $(0.05, 0.2)$ interval. Above 0.2, the density is nearly collinear with the dotted horizontal line having height 1, the benchmark uniform density. We discuss the shape of this distribution further in our structural model section and show that our model correctly captures the forces resulting in this shape.

Panel (2) shows the kernel density plot for the pure RCT balance tests. As we discuss in

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

previous sections, we expect that, 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 0.1, but other than that, the kernel density plot rises quickly to the uniform benchmark, which it tracks closely except for the highest p-values in the interval $(0.9, 1.0]$. It appears that the block of mass missing from $[0, 0.1)$ has been shifted to $(0.9, 1.0]$.

Panel (3) shows the results from 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 study flaws in this subsample and little additional removal of significant sniff tests.

Panel (4) shows kernel density plot of the pure subsample of other tests, which looks similar to that for the full sample in panel (1), which is expected because this subsample constitutes the bulk of the full sample. Below, structural estimation will allow us to distill the relative contributions of study flaws and removal from publication process from the shape of this density.

Panel (5) shows the kernel density plot 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. It is unlikely the publication process would remove these sniff tests based on their p-values. It also is unlikely the publication process would favor insignificance of pre-match sniff tests. This subsample, therefore, represents a sample of sniff tests where the alternative is true and pick up study flaws before matching is performed. We see a mountain of mass of p-values in $[0, 0.1)$, after which the density dips well below the uniform benchmark. This density does not exhibit the non-monotonicity evidenced in some preceding panels, consistent with our expectation that there is little p-value removal in this subsample. This subsample may be as close as we come to a canonical example in which the alternative assumption of misspecification holds for all studies. The shape of the density function motivates our modeling p-values under the alternative hypothesis as coming from a beta distribution.

Panel (6) shows kernel density plot of 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 line, much of this mass shifted to a spike of p-values above 0.9, suggesting matching as an overall effective method to improve balance.

4.2. Regression Specifications

In this subsection, we formalize observations we made from the kernel density plots of p-values by conducting statistical analysis on the full sample of p-values and intervals.

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

Tables differ in the precision at which they report p-values, some specifying an interval for p_i, others the exact value. To formalize these reporting conventions, let R_t denote the significance intervals reported by table t. For example, if table t does not directly report p-values, only indicating significance at the 5% level with a star, then $R_t = \{[0, 0.05), [0.05, 1], [0, 1]\}$. If table t directly reports p-values, abstracting from rounding issues, R_t equals the set of all possible subsets of $[0,1]$ (i.e., the power set, $2^{[0,1]}$). Let $\alpha \subset [0,1]$ denote the interval that we want to study. A table's reporting convention may or may not line up with α. Let $T_\alpha = \{t \in T \,|\, \alpha \in R_t\}$ denote the subpopulation of tables reporting significance in the interval α and $I_\alpha = \{i \in I \,|\, \alpha \in R_{t(i)}\}$ denote the subpopulation of sniff tests in that subpopulation of tables.

Let $S_{i\alpha} = \mathbf{1}(p_i \in \alpha)$ be the indicator of whether sniff test i is contained in interval α. We are interested in the proportion of sniff tests in the typical table falling into interval α, captured by the following conditional expectation:

$$\pi_\alpha = E_{t \in T_\alpha}(E(S_{i\alpha} \,|\, i \in t)). \tag{1}$$

This corresponds to the sampling frame that first samples tables and then sniff tests within tables. An obvious alternative is to directly sample from the population of sniff tests, leading to the expectation $E_{i \in I_\alpha}(S_{i\alpha})$. If different scholarly works contain different numbers of sniff tests and the number of sniff tests presented in the work is correlated with their significance, then the two sampling frames can generate different expectations at any arbitrary p-value cutoff between $[0,1]$, and hence different aggregate distributions. It is thus important to choose the appropriate

sampling frame and focus the analysis on the expectation associated with that frame. We choose equation (1) so each scholarly work has equal contribution to our aggregate distribution. The alternative sampling frame that collects all the sniff tests together and uses a simple unweighted aggregate causes articles with more sniff tests to be over-represented. Appendix B shows that the unweighted aggregate is a biased estimate of (1) in a predictable direction.

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

To derive the reduced-form specification used to consistently estimate π_α, we slightly abuse notation by letting T now represent the sample of tables and I the sample of sniff tests rather than the respective populations. Similarly, we redefine T_α and I_α so that they are the subsamples analogous to the subpopulations they were formerly defined as. Vertical bars denote the cardinality of sets, so $|T|$ denotes the number of tables in the sample, $|I|$ the number of sniff tests in the sample, and $|t|$ the number of sniff tests in table t.

A consistent estimator of π_α is obtained by replacing the expectations in (1) with sample averages:

$$\hat{\pi}_\alpha = \frac{1}{|T_\alpha|} \sum_{t \in T_\alpha} \left(\frac{1}{|t|} \sum_{i \in t} S_{i\alpha} \right). \tag{2}$$

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

$$w_i = \frac{|I_\alpha|}{|T_\alpha||t(i)|}. \tag{3}$$

These weights are inversely proportional to the number of sniff tests $|t(i)|$ in the including table, scaled by the constant $|I_\alpha|/|T_\alpha|$ that, as shown in Appendix B, ensures the weights to sum to $|I_\alpha|$, the number of relevant observations. Using (2) and (3), we have

$$\hat{\pi}_\alpha = \frac{1}{|I_\alpha|} \sum_{i \in I_\alpha} w_i S_{i\alpha}, \tag{4}$$

the inverse-frequency-weighted average of $S_{i\alpha}$ for the subsample of sniff tests in tables reporting interval α.[6]

Regression-based methods can be used to recover $\hat{\pi}_\alpha$.[7] In particular, in the inverse-frequency-weighted least squares (IFWLS) regression of $S_{i\alpha}$ on a constant using the subsample $i \in I_\alpha$, the coefficient on the constant term is is numerically identical to $\hat{\pi}_\alpha$ in equation (2):

$$\hat{S}_{i\alpha} = \hat{\pi}_\alpha \cdot 1 \qquad i \in I_\alpha. \tag{5}$$

Our analysis focuses on the intervals $R^* = \{[0, 0.05), [0.05, 0.1), [0.1, 0.15), [0.15, 0.2)\}$. Rather than running separate regressions for each α, the estimates can be conveniently recovered from a single regression in which multiple copies of each observation i are stacked, one copy for each interval $\alpha \in R^* \cap R_{t(i)}$ that constitutes a match between our study set and the containing table's reporting convention:

$$\hat{S}_{i\alpha} = \sum_{r \in R^*} \hat{\pi}_r \mathbf{1}(\alpha = r) \qquad \alpha \in R^*, i \in I_\alpha. \tag{6}$$

The IFWLS estimates $\hat{\pi}_\alpha$ in (5) and (6) are numerically equal, both equal to the proportion in (4). The clustered standard errors are also identical across equations (5) and (6). We cluster all standard

[6] See Wooldridge (2010), chapter 20, for a textbook discussion of the need for weighting in various sampling contexts. Equation (2) is an implementation of the estimator he suggests for cluster-sampling contexts (see his equation (20.48)). Equation (4) is an implementation of the estimator he suggests for the standard-stratified-sampling context (see his equation (20.13)). In our application, the two contexts are equivalent since, within each stratum (i.e., each table), all observations (i.e., all sniff tests) are collected.

[7] IFWLS regressions can be run in Stata using the `reg` command, setting user-defined weights (`iweights`) equal to $1/|t(i)|$. Stata's automatic scaling using this command generates weights w_i.

errors at the table (t) level; this clustering strategy correctly adjusts for the appearance of multiple copies of each observation i in the stacked regression.

The null typically used for hypothesis testing, $\pi_\alpha = 0$, is not of interest here. Under that null, there are no p-values in α, more consistent with the presence of extreme publication selection than its absence. A more natural null hypothesis for this analysis is $\pi_\alpha = |\alpha|$, where $|\alpha|$ denotes the length of interval α, corresponding with the mass of p-values that would arise under a uniform $[0,1]$ distribution, the distribution of p-values in the absence of misspecification and removal in the publication process. We will conduct separate tests of $\pi_\alpha = 0.05$ for each $\alpha \in R^*$, which each have length 0.05, as well as the joint test of $\pi_\alpha = 0.05$ for all $\alpha \in R^*$.

The joint test provides a qualitative measure of the overall departure from uniformity across thresholds $\alpha \in R^*$. To quantify the common departure from uniformity, suppose first that the p-value proportions exceed the uniform thresholds by a constant level λ: i.e., $\pi_\alpha = |\alpha| + \lambda$ for all $\alpha \in R^*$. Substituting the analogous equation involving estimators into (6) and rearranging yields $\hat{S}_{i\alpha} = (|\alpha| + \hat{\lambda}) \sum_{r \in R^*} \mathbf{1}(\alpha = r) = |\alpha| + \hat{\lambda}$. Defining $Y_{i\alpha} = S_{i\alpha} - |\alpha|$, the significance indicator normalized by its expected threshold value under a uniform distribution, and $\hat{Y}_{i\alpha} = \hat{S}_{i\alpha}$ to be its fitted value from a regression, the preceding calculations yield the following regression specification:

$$\hat{Y}_{i\alpha} = \hat{\lambda} \cdot 1 \qquad \alpha \in R^*, i \in I_\alpha. \tag{7}$$

According to this equation, an estimate of the common level departure from uniformity $\hat{\lambda}$ can be recovered from an IFWLS regression of the normalized significance indicator $Y_{i\alpha}$ on a constant; the coefficient on the constant term provides the desired estimate, $\hat{\lambda}$.

Suppose instead that the p-value proportions exceed the uniform thresholds by a constant proportion ρ: i.e., $\pi_\alpha = (1+\rho)|\alpha|$ for all $\alpha \in R^*$. Substituting the analogous equation involving estimators into (6) and rearranging yields $\hat{S}_{i\alpha} = (1+\hat{\rho})|\alpha| \sum_{r \in R^*} \mathbf{1}(\alpha = r) = (1+\hat{\rho})|\alpha|$, or upon further rearranging,

$$\hat{Y}_{i\alpha} = \hat{\rho}|\alpha| \qquad \alpha \in R^*, i \in I_\alpha. \tag{8}$$

According to this equation, an estimate of the common proportional departure from uniformity $\hat{\rho}$ can be recovered from regressing $\hat{Y}_{i\alpha}$ on $|\alpha|$ excluding a constant; the coefficient on $|\alpha|$ provides the desired estimate, $\hat{\rho}$.

4.3. Regression Results

Table 1 presents estimates $\hat{\pi}_\alpha$ of the proportion of p-values in various intervals. The estimates are computed using the stacked IFWLS regression (6). Standard error clustered at the table (t) level are reported in parentheses below the estimates. The size of the underlying subsample as well as the number of stacked observations and clusters are also reported.

Column (1) presents aggregate results for the full sample. The first entry shows that 9.8% of tests fall into the $[0, 0.05)$ interval, 3.8% fall into the $[0.05, 0.1)$ interval, and 3.7% fall into the $[0.10, 0.15)$ interval. These results are all significantly different from the 0.05 expected for the uniform benchmark. 4.7% of p-values fall into the $[0.15, 0.20)$ interval, which is not significantly different from 0.05. The pattern confirms what we observe from the kernel density plot and suggests study flaws and removal in the publication process are important forces in determining the shape of the aggregate distribution of p-values.

A cleaner test of removal alone is provided by column (2) of Table 1, which focuses on RCT balance tests, in particular the pure subsample in which balance-improving measures such as re-randomization, stratification, or matching have not been used. For this subsample, in the absence of removal, p-values should be closely resemble the uniform distribution with the percentage of p-values significant in each interval equaling the size of the interval. This is the case with the lowest interval, which has 4.1% of p-values, and the interval of $[0.15, 0.20)$, which has 5.4% of p-values, both insignificantly different from 0.05. For the other two, the percentages are less than 0.05. The estimate $\hat{\lambda} = -0.013$ implies that on average across the three low thresholds, the empirical proportions fall short of their respective benchmarks by 1.3 percentage points in level terms. The estimate $\hat{\rho} = -0.334$ implies that on average across the three low thresholds, the empirical proportions fall short of their respective benchmarks by 33.4% in proportional terms. Both estimates of common departures across thresholds are significant at the 1% level.

Column (3) shows the results from RCT balance tests when a balance improvement method (re-randomization, stratification, and/or matching) was definitively employed or likely employed. The departure from uniformity in column (3) is smaller than that in column (2), indicating either the techniques are effective in correcting for imbalances, the publication process does not differentially remove significant sniff tests in this subsample, or both.

The remaining columns (4)–(6) in Table 1 analyze tests other than RCT balance. Unlike pure

RCT balance tests, for which we expect the amount of study flaws to be small, for the "mixed bag" of other tests in columns (4)–(6), we do not know how much study flaws are present and would be picked up by sniff tests.

Shown in column (4), the results for the pure subsample of non-RCT balance tests sharply contrast those in column (2) for the pure subsample of RCT balance tests. Rather than observing the proportion of p-values below the uniform in the $[0, 0.05)$ interval as we did in column (2), we see the proportion to be far greater than the threshold value. Evidently, a substantial proportion of studies are failing sniff tests because of specification or identification problems, enough to swamp the removal process that would tend to reduce the proportion of these significant p-values.

Columns (5) and (6) present results from subsamples where a technique to improve p-values was definitively employed in the paper. Specifically, this technique is matching since, as discussed, this is the only technique of relevance for the sample of non-RCT-balance tests. Column (5) shows a large proportion of significant p-values in matched samples before matching takes place. Column (6) shows that, after matching, the proportion of significant p-values is greatly reduced, returning to values much closer to the thresholds. Matching appears to be an effective method to improve p-values in originally problematic studies, confirming the patterns observed in panels (5) and (6) of Figure 3.

5. Structural Analysis

5.1. Theoretical Background

Before presenting our structural model, we provide some basic theoretical results on mixtures distributions of p-values. We begin with the textbook result that p-values are uniformly distributed in standard settings.

Proposition 1. *The p-value of a test of a non-composite null hypothesis using a continuous test statistic is uniformly distributed on* $[0, 1]$ *under the null.*

See Lehmann and Romano (2005) for a textbook proof. The proposition places two qualifiers on the generality of the uniform distribution of p-values, supposing the test statistic is continuous and the null hypothesis is non-composite. The distribution of p-values is only approximately uniform if the test statistic is discrete (as Lehmann and Romano 2005, Example 3.3.2, illustrates).

P-values need not be uniformly distributed for tests of composite null hypotheses, i.e., nulls involving multiple distributions (Bayarri and Berger 2000).

These qualifiers are not material for our subsample of balance tests in RCTs since the difference in means between treatment and control groups is a continuous variable and the null that this difference is zero is non-composite. The vast majority of our subsample of other tests also involve continuous test statistics and non-compound nulls, being test of balance in quasi-natural experiments, tests of parallel trends in difference-in-difference designs, or tests of continuous running variables in regression-discontinuity designs. It is possible that a few of the observations in this subsample involve discrete test statistics or composite nulls, but such cases are sufficiently rare that the mixture distribution for the pooled sample discussed in the next proposition will be approximately uniform. To alleviate residual concerns, we analyze the subsample of balance in RCTs separately from other tests and keep the caveat that the distribution of p-values in the latter subsample may not be perfectly uniform under the null.

We next move from individual p-values to a pooled sample of them. The distribution of this pooled sample is the mixture distribution with probability weight w_i put on drawing sniff test i. The next proposition states that uniformity extends to the pooled sample of p-values under standard conditions.

Proposition 2. *Consider tests of n non-composite null hypotheses using continuous test statistics, which can be correlated or not. Let P be the mixture of the associated p-values from these tests, where p-value p_i is drawn with weight w_i, for $w_i \geq 0$ and $\sum_{i=1}^{n} w_i = 1$. Then P is uniformly distributed on $[0,1]$ under the null.*

Proof. By Proposition 1, each p_i has a uniform distribution on $[0,1]$. Therefore, each probability density function (pdfs) is $f_i(p) = 1$ for $p \in [0,1]$ and $f_i(p) = 0$ otherwise. Since P is a mixture of these random variables, its pdf is $f(p) = \sum_{i=1}^{n} w_i f_i(p) = 1$ for $p \in [0,1]$ and $f(p) = 0$ otherwise. □

The following proposition, due to Pounds and Morris (2003), provides theoretical underpinning for the decomposition of the aggregate p-value distribution into a mixture of null and alternative densities.

Proposition 3. *Let P be any continuous random variable with pdf $f(p)$ and support in $[0,1]$. For all $\omega \in [0,1]$ such that $\omega \leq \min\{f(p)|p \in [0,1]\}$, there exists a well-defined pdf $g(p)$ such that $f(p) = \omega + (1-\omega)g(p)$.*

Combining the last two propositions, a mixture distribution of p-values from a set of sniff tests can be decomposed into two components: one arising from sniff tests of well-specified studies for which the null is true and one arising from sniff tests of misspecified studies for which the alternative is true. By Proposition 2, the distribution for tests under true nulls is uniform on $[0,1]$, extracted as a unit density weighted by ω in Proposition 3. We will flexibly approximate the alternative density $\tilde{f}$ by a beta distribution, used elsewhere in the literature, and estimate the proportions of well-specified and misspecified studies, ω and $1-\omega$, in the sample.

5.2. Model

Our structural model posits a two-stage process for generating p-values. In the first stage, p-values are drawn from two populations of studies: a proportion ω of studies for which the null hypothesis of no flaws is true and the complementary proportion, $1-\omega$, of studies exhibiting flaws for which the alternative hypothesis is true. Building on Sellke, Bayarri, and Berger's (2001) suggestion of using a beta distribution with one free parameter as the alternative to a uniform distribution of p-values, we assume that the initial draw of p-values from the alternative population follows a the beta distribution with parameters $1/(1+\mu)$ and 1. The associated density is

$$g(p_i;\mu) = \left[(1+\mu)p_i^{\frac{\mu}{1+\mu}}\right]^{-1}. \tag{9}$$

The shape parameter $\mu \in [0,\infty)$ reflects the severity of misspecfication. When $\mu = 0$, there are no flaws in the population of studies, and g reduces to the uniform density. When μ increases, g becomes increasingly left-skewed, piling up mass on the lowest p-values.

The first-stage density function is a mixture of a proportion ω of well-specified studies with a uniform distribution of p-values and $1-\omega$ misspecified studies with p-values given by density g:

$$f_1(p_i;\mu,\omega) = \omega + (1-\omega)g(p_i;\mu). \tag{10}$$

In the second stage, the publication process results in the removal of some sniff tests with significant p-values. For simplicity, start with the case of a homogeneous removal rate $\rho \in [0,1]$ in a single removal region $\alpha = [0,\hat{\alpha})$. Below we will generalize to multiple regions with heterogeneous

removal rates. Applying Bayes Rule to the prior density in equation (10), the posterior density emerging from the second stage for p-values in our sample is

$$f(p_i;\mu,\omega,\rho) = \frac{(1-S_{i\alpha}\rho)f_1(p_i;\mu,\omega)}{1-\rho\left[\omega\hat{\alpha}+(1-\omega)\hat{\alpha}^{\frac{1}{1+\mu}}\right]}. \tag{11}$$

We will call this the beta-uniform mixture (BUM) model, in particular, the variant with a single removal region, BUM(SR). Appendix B derives the density function for the extension of the BUM model to multiple removal regions $\alpha_k = [\hat{\alpha}_{k-1}, \hat{\alpha}_k)$ with different removal rates ρ_k in each, labeled the BUM(MR) model. We will call the variant of the BUM model in which there is no removal in the publication process BUM(NR). We will call the special case in which there are only misspecified studies—with no mixture of uniformly distributed well-specified studies—the Beta model. The variants with either no removal, a single removal region, or multiple removal regions are labeled Beta(NR), Beta(SR), and Beta(MR), respectively. The density function can be obtained from (11) by substituting $\omega = 0$.

Some technical remarks about the BUM model are in order. First, we labeled μ as the severity of misspecification. More precisely, it reflects the combination of latent misspecification and the power of the sniff tests used in our sample studies to detect this latent misspecification. Our model does not separately identify these two components of measured misspecification. Second, the model assumes no removal outside of α. In reality, studies are removed in the publication process for reasons unrelated to the performance of sniff tests that our model does not capture. Hence, ρ is best interpreted as the marginal increase in the removal rate when the p-value from a sniff test is in interval α. This interpretation accommodates the possibility that p-value p_i outside of α is removed as a side effect of other sniff tests in the same table falling into α. Third, the model assumes the same ρ applies to well-specified and misspecified studies. In effect, participants in the publication process do not have additional knowledge about the quality of studies' specification beyond the significance of sniff-test results.

Figure 4 illustrates the shape of the BUM(SR) density for various parameters. Panel (1) graphs the density when there is removal (various removal rates ρ shown) but no misspecification, i.e., $\mu = 0$. The removal interval is set to $\alpha = [0, 0.15)$, consistent with patterns in reduced-form analysis. These densities have less mass on p-values below 0.15 and more mass above 0.15 than does the

uniform distribution. The larger is ρ, the lower the mass in $[0, 0.15)$. Panel (2) graphs the density in the absence of removal ($\rho = 0$) but in the presence of study flaws (various values of μ shown). For illustrative purpose, the proportion of well-specified studies is set to $\omega = 2/3$. These densities become increasingly left-skewed for higher μ. The last two panels graph BUM(SR) densities when both removal and study flaws are present. Panel (3) varies μ holding ω constant; panel (4) varies ω holding μ constant. Except for the degenerate cases of $\mu = 0$ or $\omega = 1$ (parameter restrictions that eliminate misspecified studies), the twin forces of removal and misspecifiation lead the densities to be non-monotonic, starting from a spike at the lowest p-values, declining over the interval $(0, 0.15)$ until they fall below the uniform benchmark, then jumping above the uniform benchmark at a p-value of 0.15, and declining from there to the end of the unit interval.

The densities from the BUM(SR) model in Figure 4 closely resemble the empirical densities in Figure 3. Panel (1) of Figure 4, with removal but no study flaws, resembles panels (2) and (3) of Figure 3, from subsamples of balance tests in RCTs, for which one should expect few study flaws in the form of randomization failures. Panel (2) of Figure 4, with no removal but with study flaws, resembles panel (5) of Figure 3, from a subsample that is sufficiently flawed for authors to take actions to improve p-values. Since removal is likely based on the post-improvement rather than the pre-improvement p-values, it would not be surprising for this subsample to exhibit little removal. The last two panels of Figure 4, which exhibit both removal and misspecification, closely resemble panels (1) and (4) of Figure 3 for samples that likely were exposed to those two forces. The fact that the density from the model follows the same unusual non-monotonic pattern as these empirical densities provides some confidence in the validity of our structural model.

5.3. Estimation

We will use the method of maximum likelihood to estimate the parameters of the BUM model. The density function (11) provides the building block for the likelihood function. To derive the log-likelihood function, we start with the simple case of a single removal region $\alpha = [0, \hat{\alpha})$ with removal rate ρ.

For any subsample under consideration, some of the observations i will have exact p-values p_i reported; for others all we can glean is the interval $[\ell_i, u_i)$ containing p_i. The following log-

likelihood function includes components for both sorts of observations:

$$\ln L = \sum_{p_i \text{ exact}} \ln f(p_i;\mu,\omega,\rho) + \sum_{p_i\in[\ell_i,u_i)} \ln \int_{\ell_i}^{u_i} f(p;\mu,\omega,\rho)\,dp \tag{12}$$

$$\begin{aligned} &= N_{p_i<\hat{\alpha}} \ln(1-\rho) - N \ln\left(1-\rho\left[\omega\hat{\alpha} + (1-\omega)\hat{\alpha}^{\frac{1}{1+\mu}}\right]\right) + \sum_{p_i \text{ exact}} \ln\left(\omega + \frac{1-\omega}{1+\mu} p_i^{\frac{-\mu}{1+\mu}}\right) \\ &\quad + \sum_{p_i\in[\ell_i,u_i)} \ln\Bigg(\omega\left\{(1-\rho)\left[\min(u_i,\hat{\alpha}) - \min(\ell_i,\hat{\alpha})\right] + \max(u_i,\hat{\alpha}) - \max(\ell_i,\hat{\alpha})\right\} \\ &\qquad + (1-\omega)\Bigg\{(1-\rho)\left[\min(u_i,\hat{\alpha})^{\frac{1}{1+\mu}} - \min(\ell_i,\hat{\alpha})^{\frac{1}{1+\mu}}\right] \\ &\qquad\qquad + \max(u_i,\hat{\alpha})^{\frac{1}{1+\mu}} - \max(\ell_i,\hat{\alpha})^{\frac{1}{1+\mu}}\Bigg\}\Bigg), \end{aligned} \tag{13}$$

where N denotes the number of observations in the subsample under consideration and $N_{p_i<\hat{\alpha}}$ denotes the number of those reporting an exact p-value that is less than $\hat{\alpha}$.

Maximum likelihood estimates, which will be denoted with tildes, $\tilde{\mu}$, $\tilde{\omega}$, and $\tilde{\rho}$, can be obtained by maximizing $\ln L$. The inverse-frequency weights w_i used in the reduced-form analysis will also be applied here in the maximum-likelihood procedure; see Section 4.1 for a discussion of the rationale.

Appendix B provides the log-likelihood in the general case of the BUM model with multiple removal regions. We will estimate the BUM(MR) model allowing for different removal rates $\tilde{\rho}_1$, $\tilde{\rho}_2$, $\tilde{\rho}_3$, and $\tilde{\rho}_4$ in the intervals $[0,0.05)$, $[0.05,0.10)$, $[0.10,0.15)$, and $[0.15,0.20)$, respectively. To economize on parameters, we also estimate the BUM(SR) variant with a single removal region. Based on visual inspection of the kernel densities in Figure 3 and reduced-form regression results in Table 1, we take the single removal region to be $\hat{\alpha} = [0,0.15)$ in that case.

5.4. Structural Results

The structural estimates are reported in Table 2. Column (1) reports results for the pure subsample of RCT balance tests. RCT balance tests are good candidates for focused study for several reasons. Since this is a test is of a non-compound hypothesis using a continuous statistic, the distribution of p-values in the absence of removal and misspecification is uniform by Proposition 1. Furthermore, even allowing for misspecification, the mixture distribution should be close to uniform since

misspecification comes in the form of randomization failure and such failures are likely to be rare. Column (1) reports the most general BUM(MR) specification; columns (2) and (3) report more restrictive specifications for the same subsample.

In column (1), we estimate a proportion misspecified of $1-\tilde{\omega}=0.013$, not significantly different from zero. The small number of misspecified studies (1.3% in percentage terms) is expected given the rarity of randomization failures, the source of misspecification in this subsample. The few misspecified studies are severely misspecified, as we estimate $\tilde{\mu}=10.179$, signficantly different from 0 at the one-percent level. The removal rates $\tilde{\rho}_1$, $\tilde{\rho}_2$, and $\tilde{\rho}_3$ in the intervals $[0,0.05)$, $[0.05,0.10)$, and $[0.10,0.15)$ are significantly different from 0 at the 1% level, ranging from 32.2% to 58.5% in percentage terms. By comparison, the removal rate $\tilde{\rho}_4$ estimated for the interval $[0.15,0.20)$ is small at 3.2% and not statistically significant.

Column (2) presents results for the same sample but for the BUM(SR) model restricted to a single removal region. Based on earlier reduced-form evidence as well as the structural evidence here that removal dies out above 0.15, we take the removal region to be $[0,0.15)$. A likelihood-ratio test cannot reject the BUM(SR) in favor of the BUM(MR) model; though the BUM(MR) involves three more parameters, the two models' log-likelihoods are nearly identical. The estimates of $1-\tilde{\omega}$ and $\tilde{\mu}$ are similar in column (2) to column (1). The removal rate $\tilde{\rho}$ over the interval $[0,0.15)$ is an estimated 45.7%.

Column (3) further restricts the model to involve just the beta alternative, eliminates the parameter ω allowing for a mixture of well-specified studies. Since there is so little misspecification in this subsample, the model adapts by reducing $\tilde{\mu}$ to close to 0 and estimating what is nearly a uniform distribution; the removal rate is reduced to accommodate the uniform approximation. The p-value for the likelihood-ratio test of Beta(SR) against BUM(SR) is 0.17, indicating that the fit of the Beta is not substantially worse for this subsample. We will see below a subsample with more misspecification for which the restriction from BUM to Beta is strongly rejected.

Columns (4)–(6) report results for the pure sample of tests other than balance in RCTs. A handful of these tests may involve compound hypotheses or may be tested with discrete statistics, so the distribution of p-values may not be exactly uniform. However, the vast majority of these tests satisfy the conditions from Proposition 1 for uniformity, so any departure from uniformity absent removal and misspecification should be small. The subsample in columns (4)–(6) still focuses on

pure tests, excluding observations where matching was involved.[8] We use the same convention as the other subsample, starting with the most general BUM(MR) specification in column (4) and moving to more restricted specifications in columns (5) and (6).

For other tests, we obtain larger values for $1-\tilde{\omega}$ than for tests of balance in RCTs: 9.0% in the BUM(MR) and 9.4% in the BUM(SR) model. Both are significantly different from zero at the one-percent level, suggesting a significant proportion of flawed studies in this subsample. The estimates $\tilde{\mu}=9.505$ and $\tilde{\mu}=9.225$ in columns (4) and (5) are similar to those in columns (1) and (2), suggesting similar shapes of alternative densities in the two subsamples. This suggests that the severity of study flaws and the statistical power of sniff tests in picking up study flaws are similar across the two samples. As with the subsample of tests of balance in RCTs, removal rates $\tilde{\rho}_1$–$\tilde{\rho}_3$ in the lowest three intervals are statistically significant and fairly homogenous, but the removal rate $\tilde{\rho}_4$ is close to 0 (indeed negative) and not statistically significant. The main difference is that the removal rate is only about half that for the previous subsample. In column (5), which estimates a single removal rate over the interval $[0,0.15)$, the estimate of $\tilde{\rho}$ is 25.1%, compared to 45.7% in column (2).

The further restriction from BUM(SR) to Beta(SR) for the subsample of other tests turns out to be rejected at the 1% level. The mixture fits much better than the single beta distribution. Across the two samples, BUM(SR) with the removal region set to $[0,0.15)$ fits the data relatively well and parsimoniously.

Table 3 leverages the structural estimates to provide insight into whether the publication process leads to too much or too little removal. Each panel is a contingency table, with boxed cells showing the percentage of the subsample falling into that cell. Percentages in the boxed cells thus sum to 100%. Panel A presents results for the subsample of pure RCT balance tests, panel B presents results for pure other tests, and panel C presents formulas used to compute entries from the structural estimates.

For the subsample of pure balance tests, panel A shows that a policy of refraining from remov-

[8] We excluded matched samples for several reasons. First, all the articles eventually substitute post-match for pre-match data, so the pre-match p-values do not bear on data used for actual analysis in any article. Second, as shown in Figure 3, the distribution of p-values for the pre-match subsample is wildly different from any other subsample. Although the pre-match subsample is relatively small, including it might add undue noise, obscuring any clear findings. Having excluded the pre-match subsample, it seemed natural to focus on a pure sample by excluding any matching entirely.

ing studies with p-values in the $[0.15, 1]$ region does little harm since very few misspecified studies get through. Only $0.5\%/83.5\% = 0.6\%$ of the studies are misspecified according to the structural estimates. But the rate of misspecification is so low that even with a greater concentration in the $[0, 0.15)$ region, only $1.8\%/16.5\% = 10.9\%$ of these are misspecified. The remaining 89.1% are well-specified but happened to have an unlucky sniff-test result. If, as the structural model assumes, the publication process cannot aim well enough to target misspecified studies but just randomly removes them, the literature loses about nine well-specified studies for every misspecified study removed in the $[0, 0.15)$ region. The cost of misspecification would have to be enormous for this tradeoff to be worthwhile. Recalling that the estimated removal rate is $\tilde{\rho} = 46.7\%$, if removal is random rather than targeting misspecified studies, $46.7\% \times 14.7\% = 6.7\%$ of all studies conducted are well-specified but lost to the literature (or at least this set of journals). Even if no all misspecified studies are missed by removal, the excess of the estimated removal rate $\tilde{\rho} = 45.7\%$ over the rate of misspecification (10.9%) in the $[0, 0.15)$ region means that more than three well-specified studies are lost for each misspecified study removed in that region.

In the subsample of pure other tests analyzed in panel B, the concentration of misspecification in low p-values means that, again, few studies with p-values in the $[0.15, 1]$ interval are misspecified, only $1.6\%/78.6\% = 2.0\%$ of them. In the $[0, 0.15)$ interval, $7.8\%/21.4\% = 36.4\%$ of studies are estimated to be misspecified. This exceeds the removal rate, estimated to be $\tilde{\rho} = 25.1\%$. Whether this is the efficient removal rate depends on how accurately misspecification can be targeted and the cost of allowing misspecified studies through versus the loss of well-specificed research. There are plausible parameters for which the removal rate could be efficient. If no misspecified studies are missed for removal, the removal rate is insufficient, undershooting by $36.4\% - 25.1\% = 11.3$ percentage points. However, if removal is not this accurate, and loss of well-specified studies sufficiently costly, the removal rate could be justified.

5.5. Author Claims

Authors have incentives to attribute unfavorable sniff tests to random bad luck. Such claims are difficult to dispute on an individual basis. In this subsection, we investigate whether authors—in the aggregate—tend to over- or under-attribute unfavorable sniff tests to bad luck. To do so, we combine our structural estimates of latent proportion of well-specified 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", "no claim" and "admit rejected." The "strong claim" category includes cases in which the authors express satisfaction with the test outcome. Authors express satisfaction—with good reason—when the associated p-value under consideration is insignificant. We also consider authors to be strongly satisfied whenever, faced with having to explain a significant p-value, they explicitly attribute it to random chance rather than some systematic feature of the data. These cases are often associated with tables reporting multiple sniff tests, only a few of which are significant. The "weak claim" category includes cases in which, faced with some significant sniff-test results to explain, perhaps too many to attribute to random chance, the authors are forced to acknowledge possible problems 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 effects but then argue that it does not undermine the validity of their main results, often using the argument that the significant sniff-test results follow no systematic patterns. When the authors do not discuss the specific test statistic, we classify it as "no claim." "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. The majority of observations (71%) are classified as "strong claims," reflecting in part the frequency of tables lacking problematic p-values to begin with. "Weak claims" account for 9% of observations, "no claims" for 7%, and "admit rejected" for 13% of observations.

Figure 5 compares the proportion of claims that are strong for each p-value (estimated via local polynomial smoothing) to the latent proportion of well-specified studies derived from our structural estimates. Separate panels are provided for the pure subsamples of RCT balance tests and other tests. The figure suggests that authors tend to be too conservative in attributing significant p-values to random bad luck. Except for a tiny region of p-values extremely close to 0, the proportion of p-values with strong claims is less than the proportion of well-specified studies implied by our structural estimates. The gap between the two proportions is quite large for most of the p-values displayed (for the interval $[0, 0.2]$). Whether self-imposed or forced on authors by reviewers, excess conservatism in claiming bad luck may be a symptom of the same force that leads the publication

process to reject many well-specified studies.

6. Conclusion

This paper shifts the focus of the typical meta-analysis from standard statistical tests of results of main interest—for which authors prefer rejecting null results—to ancillary tests, dubbed "sniff tests"—for which authors prefer not to reject null results. We investigate the extent to which the publication process uses sniff tests to screen out studies with questionble credibility rather than just providing additional information that readers could to judge the credibility of published studies. If they are used as a screen, we would expect sniff tests to exhibit too few significant p-values rather than too many as in traditional studies of publication bias. Besides providing a meta-analysis of a new type of test, a contribution of our paper is the painstaking collection of nearly 30,000 sniff tests from 892 articles published in 59 economics journals, which we analyze.

Certain subsamples are sufficiently well-behaved that a missing mass of significant p-values removed by the publication process can be identified from reduced-form methods alone. This is the case in particular for the subsample of RCT balance tests that are pure in the sense of coming from contexts precluding authors from applying balance-improvement techniques. Given that the only source of misspecification for this subsample is ruined randomization, which should be expected to be rare, the latent distribution of sniff-test p-values should be approximately uniform in the absence of removal of significant p-values by the publication process. Instead, there appears to be a block of missing mass in the $[0, 0.15)$ interval for this subsample. The picture for other subsamples is less clean since an unknown proportion of misspecified studies may skew the distribution of sniff-test p-values toward 0, filling in mass removed by the publication process. Still, the kernel density for other relevant subsamples exhibit a non-monotonic pattern conforming to our structural model of a mixture of well-specified and misspecified studies, some of which are removed by the publication process if their sniff-test p-values fall into certain intervals.

Bolstered by the fit between the structural model and the kernel densities for various subsamples, we proceed to estimate the parameters of the structrual model via maximum likelihood. We find that a beta-uniform mixture model with a single removal region, the BUM(SR) model, fits the data well and parsimoniously. For the pure sample of balance tests in RCTs, estimates from

this model suggest that 46% of sniff tests with p-values in $[0, 0.15)$ are removed by the publication process, while only about one in ten of them is misspecified. The rest have significant p-values not due to randomization failure but due to random bad luck in the outcome of tests on the chosen covariates. Instead of being screened out of the journals, the well-specified studies could have been resuscitated by simply conditioning on the "offending" covariates. Given there was no real randomization failure, there is no extra danger of correlation of outcomes with omitted covariates polluting estimates of treatment effects.

For the pure sample of other tests, misspecification can arise from a wide variety of sources, so the latent rate is likely much higher than with balance tests in RCTs. Indeed, we estimate the rate of misspecification is nearly seven times higher for other tests than for balance tests in RCTs. Yet the removal rate—around 25% in the $[0, 0.15)$ region of p-values in the BUM(SR) model—is only about half that for balance tests in RCTs.

Comparing authors' qualitative characterization of their own sniff tests to structural estimates of the proportion of well-specified studies, we find that authors under-attribute significant sniff tests to bad luck. Whether their own reluctance or imposed by reviewers, this finding appears to be a symptom of the same force that leads the publication process to reject many well-specified studies.

We do not perform a formal welfare analysis, lacking information on the cost of screening out a well-specified study versus the cost of publishing a misspecified one. We can conclude that if other tests besides balance in RCTs have been used as an efficient screen by the publication process, and if the welfare tradeoffs are similar for other tests as for balance in RCTs, our results suggest that too many RCTs are screened on the basis of significant results in balance tests. In any event, our paper provides information that can help editors refine their use of sniff tests to screen papers, up to now informed by anecdotal experience or gut feelings rather than a large sample of tests and formal statistical analysis.

References

2015 Journal Citation Reports. Thomson Reuters, 2015.

Allison, David. B., Xiangqin Cui, Grier P. Page, and Mahyar Sabripour. (2006) "Microarray Data Analysis: From Disarray to Consolidation and Consensus," *Nature Reviews Genetics* 7: 55–65.

Andrews, Isaiah, Matthew Gentzkow, and Jesse M. Shapiro. "On the Informativeness of Descriptive Statistics for Structural Estimates," National Bureau of Economic Research working paper no. 25217.

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

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

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

Bayarri, M. J. and James O. Berger. (2000) "*P* Values for Composite Null Models," *Journal of the American Statistical Association* 95: 1127–1142.

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

Borusyak, Kirill and Xavier Jaravel. (2017) "Revisiting Event Study Designs." SSRN working paper no. 2826228.

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

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

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

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

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

Datta, Susmita and Somnath Datta. (2005) "Empirical Bayes Screening of Many P-Values with Applications to Microarray Studies," *Bioinformatics* 21: 1987–1994.

Do, Kim-Anh, Peter Ḿuller, and Feng Tang. (2005). "A Bayesian Mixture Model for Differential Gene Expression," *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54: 627–644.

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

Freyaldenhoven, Simon, Christian Hansen, and Jesse M. Shapiro. (2018). "Pre-Event Trends in the Panel Event-Study Design," National Bureau of Economic Research working paper no. w24565.

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

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

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

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

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

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

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

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

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

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

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

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

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

Roth, Jonathan. (2020). "Pre-test with Caution Event-study Estimates After Testing for Parallel Trends," Harvard University working paper.

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

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

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

Table 1: Proportion of Significant P-values in Various Subsamples

		RCT balance tests		Other tests		
	Full sample	Pure	Possibly improved	Pure	Matched, pre-match	Matched, post-match
	(1)	(2)	(3)	(4)	(5)	(6)
Proportion of p-values in different intervals, $\hat{\pi}_\alpha$						
• $[0, 0.05)$	0.098^{***}	0.041	0.056	0.091^{***}	0.551^{***}	0.041
	(0.005)	(0.007)	(0.010)	(0.005)	(0.033)	(0.009)
• $[0.05, 0.10)$	0.038^{***}	0.025^{***}	0.037^{***}	0.039^{***}	0.070^{*}	0.024^{***}
	(0.002)	(0.005)	(0.005)	(0.003)	(0.011)	(0.005)
• $[0.10, 0.15)$	0.037^{***}	0.033^{***}	0.039^{**}	0.039^{***}	0.030^{***}	0.031^{**}
	(0.003)	(0.006)	(0.005)	(0.004)	(0.006)	(0.009)
• $[0.15, 0.20)$	0.047	0.054	0.038^{***}	0.051	0.036^{**}	0.044
	(0.004)	(0.009)	(0.004)	(0.007)	(0.006)	(0.008)
• Joint F-test	42.3^{***}	10.1^{***}	5.0^{***}	20.5^{***}	69.4^{***}	8.3^{***}
Combining results across $[0, 0.15)$						
• Linear departure from uniform, $\hat{\lambda}$	0.012^{***}	-0.013^{***}	-0.004	0.012^{***}	0.154^{***}	-0.014^{***}
	(0.002)	(0.003)	(0.004)	(0.003)	(0.011)	(0.004)
• Proportional departure from uniform, $\hat{\rho}$	0.274^{***}	-0.334^{***}	-0.109	0.262^{***}	3.89^{***}	-0.349^{***}
	(0.047)	(0.076)	(0.089)	(0.060)	(0.273)	(0.105)
Observation counts						
• Sample size	28,798	2,953	4,023	16,026	2,372	3,424
• Stacked observations	78,719	8,265	13,347	39,224	7,537	10,346
• Clusters	1,313	101	138	922	99	152

Notes: Results are the weighted proportions $\hat{\pi}_\alpha$ of observations i that have p-values in the interval α in the row heading. Weights equal w_i. Results can equivalently be obtained as coefficients from the stacked IFWLS regression (6). Each sniff-test statistic i in the sample may contribute several observations to the stacked regression, depending on the number of intervals under study $\alpha \in R^* = \{[0, 0.05), [0.05, 0.10), [0.10, 0.15), [0.15, 0.20)\}$. Column (3) combines possibly pure and definitively improved samples. Standard errors, reported in parentheses below results, are clustered at the table level. Clustering correctly adjusts standard errors for having multiple intervals stacked per observation. Significantly different from 0.05 in a two-tailed test at the *ten-percent level, **five-percent level, ***one-percent level.

Table 2: Structural Estimates

	RCT balance tests, pure			Other tests, pure		
	BUM(MR)	BUM(SR)	Beta(SR)	BUM(MR)	BUM(SR)	Beta(SR)
	(1)	(2)	(3)	(4)	(5)	(6)
• Proportion misspecified, $1-\omega$	0.013	0.023		0.090***	0.094***	
	(0.008)	(0.016)		(0.026)	(0.020)	
• Severity of misspecification, μ	10.179***	7.192**	0.050	9.505***	9.225***	0.650***
	(3.733)	(2.822)	(0.082)	(3.563)	(3.096)	(0.078)
• Removal parameters:						
◦ Removal in $[0,0.05)$, ρ_1	0.322**			0.225**		
	(0.128)			(0.107)		
◦ Removal in $[0.05,0.10)$, ρ_2	0.585***			0.261***		
	(0.076)			(0.066)		
◦ Removal in $[0.10,0.15)$, ρ_3	0.355***			0.209**		
	(0.124)			(0.089)		
◦ Removal in $[0.15,0.20)$, ρ_4	0.032			−0.027		
	(0.165)			(0.150)		
◦ Removal in $[0,0.15)$, ρ		0.457***	0.351***		0.251***	0.391***
		(0.065)	(0.064)		(0.061)	(0.023)
Log-likelihood	−12.48	−12.56	−13.50	−349.81	−349.78	−384.95
Observations	2,974	2,974	2,974	16,528	16,528	16,528
Clusters	103	103	103	951	951	951

Notes: Results from inverse-frequency-weighted maximization likelihood (IFWML). Weights equal w_i. For the small number of observations whose p-values are smaller than machine double precision, producing an undefined natural logarithm, we rounded up to machine double precision; results nearly identical if these few observations are dropped. We report proportion misspecified $1-\omega$ rather than proportion well-specified ω to facilitating testing against natural null hypothesis that estimate equals 0. Standard errors clustered at the table t level reported in parentheses. Stars indicate significant difference from 0 in a two-tailed test at the *ten-percent level, **five-percent level, ***one-percent level.

Table 3: Latent Misspecification Prior to Removal Derived from Structural Estimates

A. RCT balance tests, pure

	$p_i \in [0, 0.15)$	$p_i \in [0.15, 1]$	Row total
Misspecified studies	1.8%	0.5%	2.3%
Well-specified studies	14.7%	83.0%	97.7%
Column total	16.5%	83.5%	100%

B. Other tests, pure

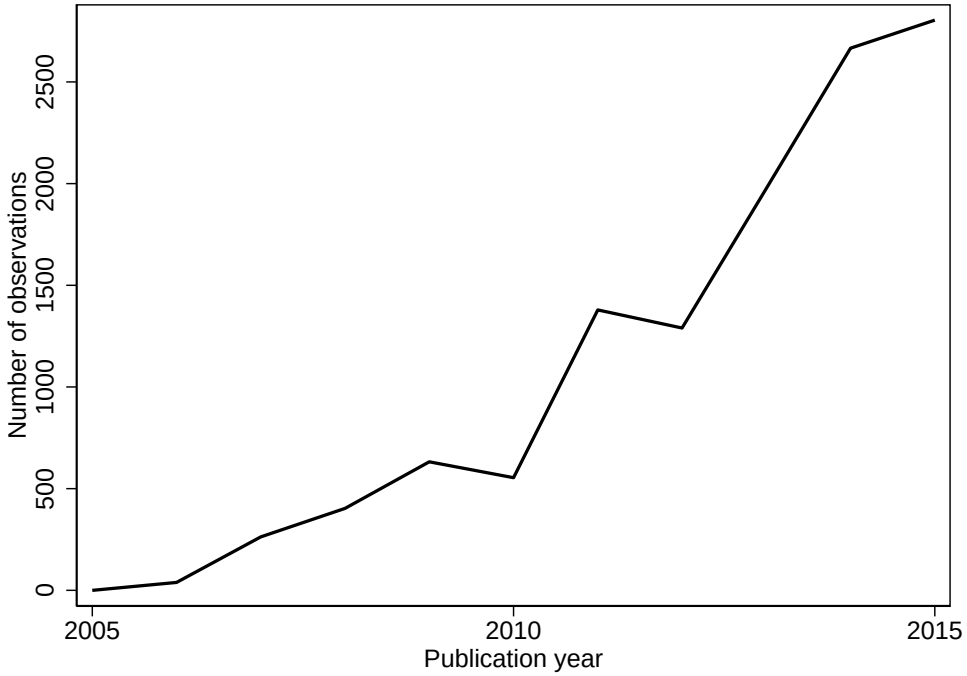
	$p_i \in [0, 0.15)$	$p_i \in [0.15, 1]$	Row total
Misspecified studies	7.8%	1.6%	9.4%
Well-specified studies	13.6%	77.0%	90.6%
Column total	21.4%	78.6%	100%

C. Formulas

	$p_i \in [0, \hat{\alpha})$	$p_i \in [\hat{\alpha}, 1]$	Row total
Misspecified studies	$(1-\tilde{\omega})\int_0^{\hat{\alpha}} g(p;\tilde{\mu})dp$	$(1-\tilde{\omega})\int_{\hat{\alpha}}^1 g(p;\tilde{\mu})dp$	$1-\tilde{\omega}$
Well-specified studies	$\tilde{\omega}\hat{\alpha}$	$\tilde{\omega}(1-\hat{\alpha})$	$\tilde{\omega}$
Column total	$\int_0^{\hat{\alpha}} f_1(p;\tilde{\mu},\tilde{\omega})dp$	$\int_{\hat{\alpha}}^1 f_1(p;\tilde{\mu},\tilde{\omega})dp$	100%

Notes: Panels A and B are contingency tables for the indicated subsamples. Boxed cells report percentage of the subsample in that cell. Entries in boxed cells sum to 100% within each panel. Panel C provides formulas for computing entries from structural parameters estimated for the BUM(SR) model with removal region $\alpha = [0, \hat{\alpha})$. In the first two panels, $\hat{\alpha}$ is set to 0.15.

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

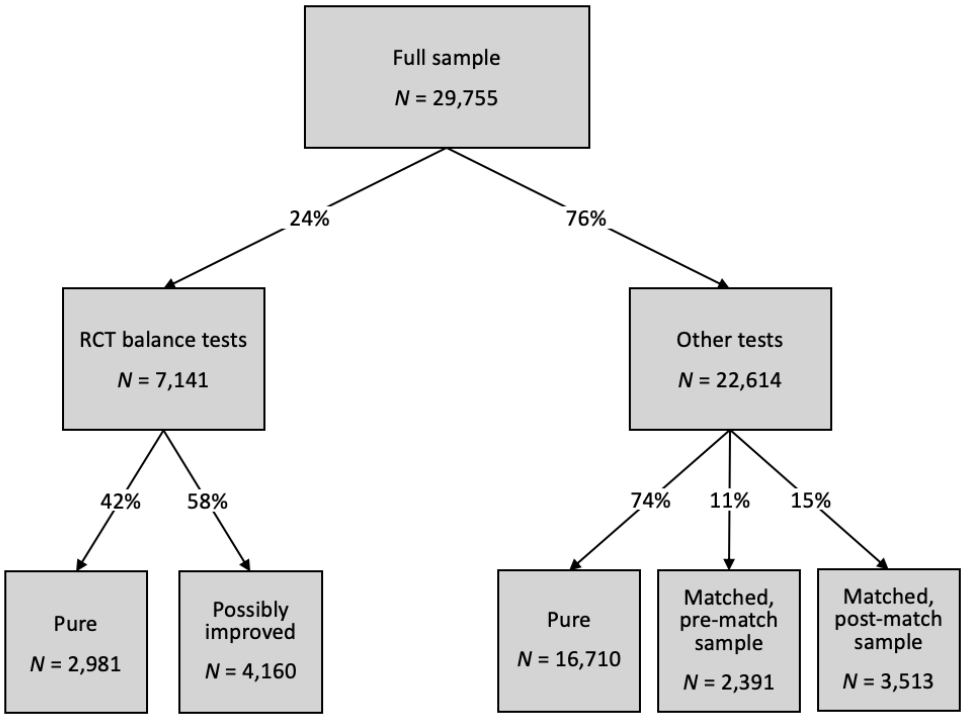
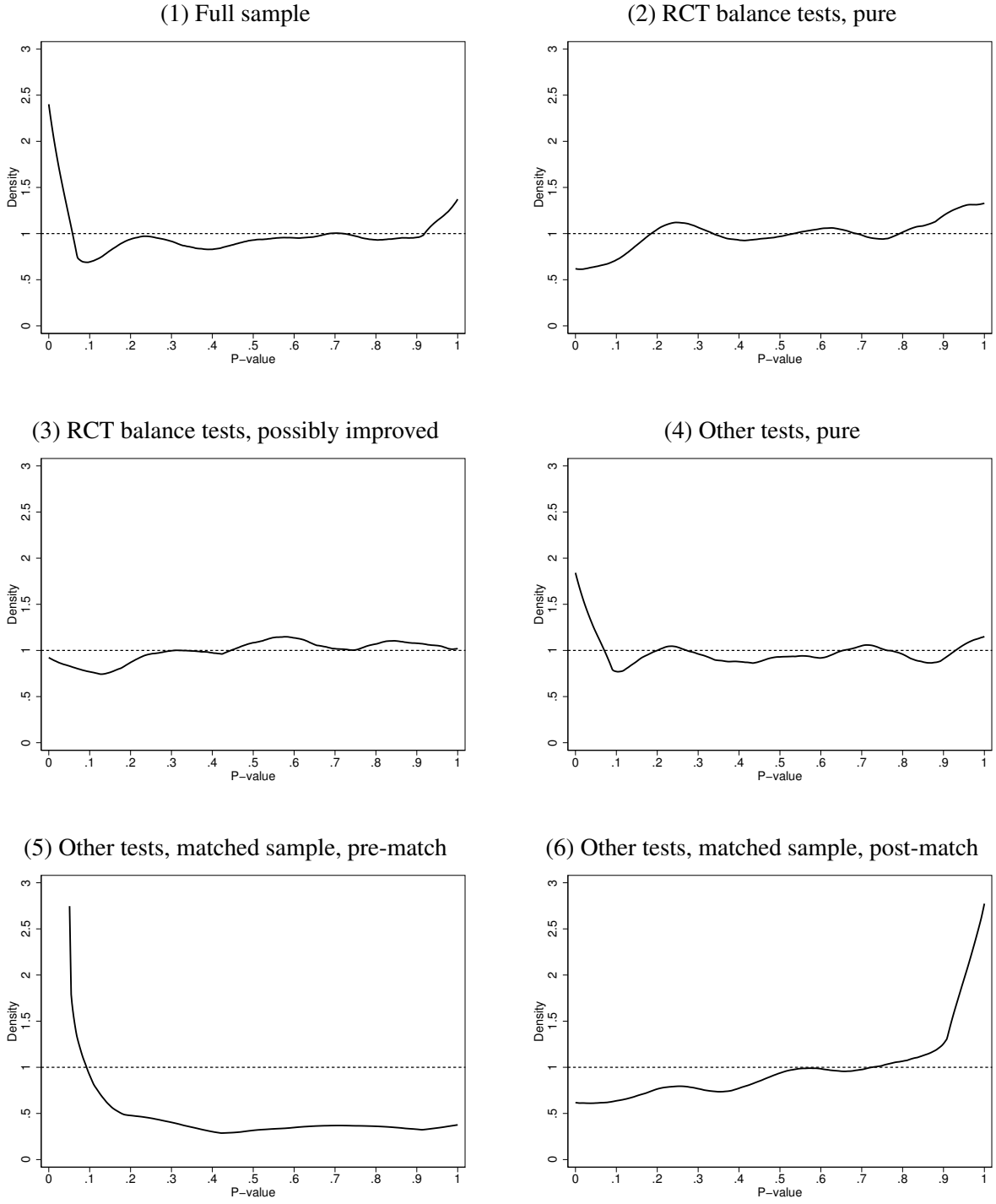
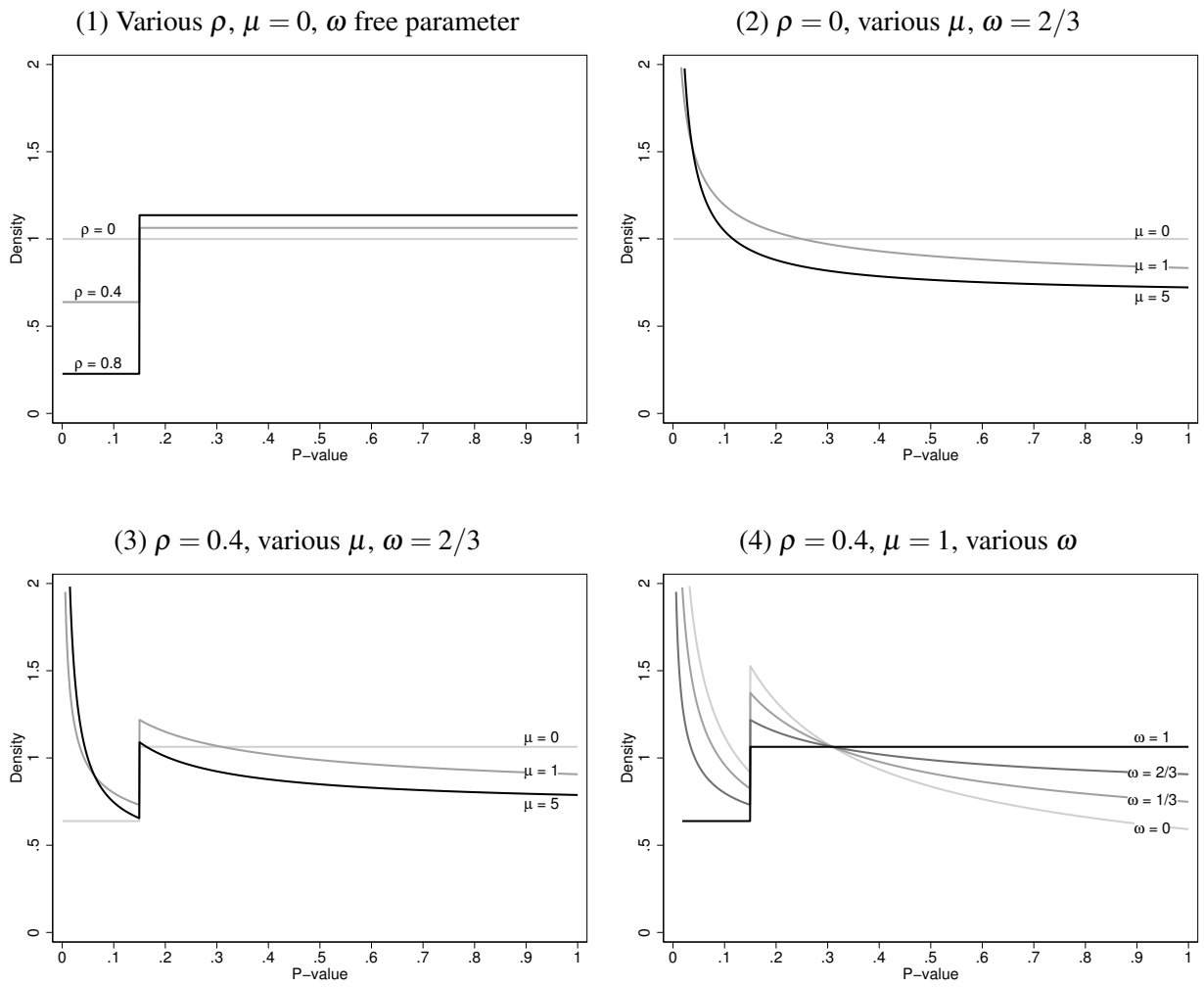


Figure 3: Kernel Density Estimates of Distribution of P-values



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

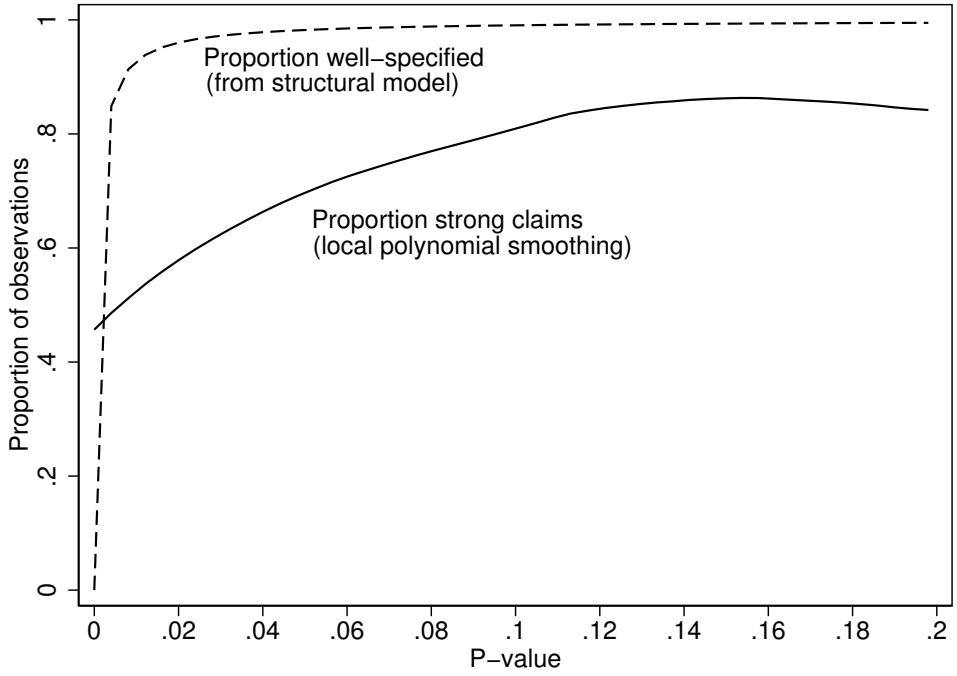
Figure 4: P-value Densities Generated by BUM Model



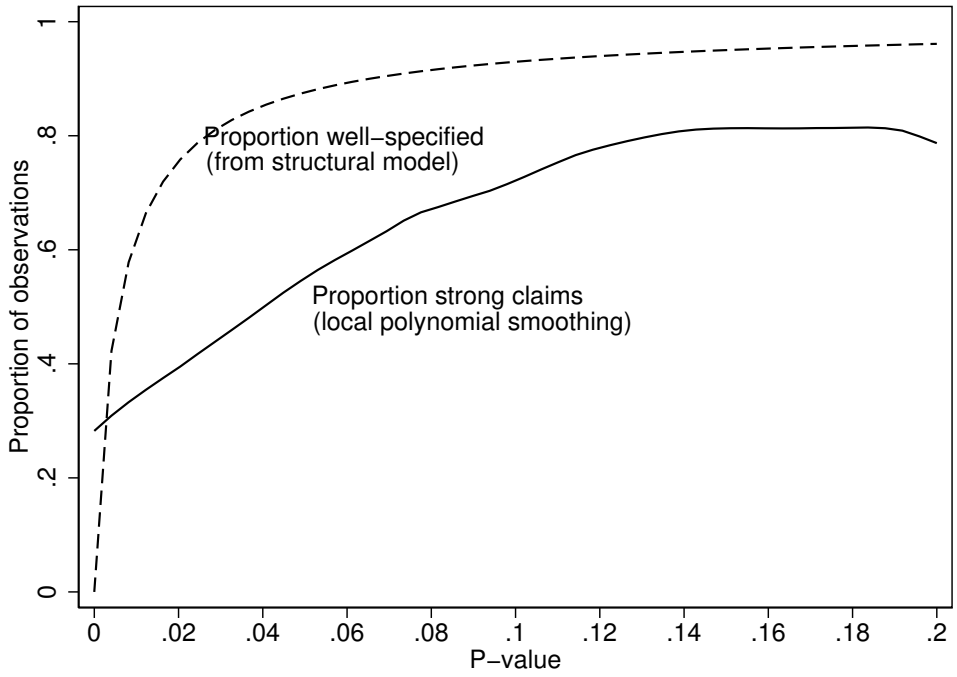
Notes: Plots of density $f(p;\mu,\omega,\rho)$ from BUM model for single removal interval $\alpha = [0,0.15)$ and for indicated parameters. With $\mu = 0$ in panel (1), ω does not affect the shape of the curves. The panel would be the same setting $\omega = 1$ and letting μ be a free parameter. Several curves represent degenerate cases: the $\rho = 0$ curve in panel (1) (in which $\mu = 0$) and the $\mu = 0$ curve in panel (2) (in which $\rho = 0$) are equivalent to the uniform density; the $\omega = 0$ curve in panel (4) is equivalent to the Beta model.

Figure 5: Strength of Author Claims

(1) RCT balance tests, pure



(2) Other tests, pure



Note: Figures show the proportion of well specified test statistics according to the authors of papers in our sample and the structural estimates from Table 2 columns (2) and (5).

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

Impact Factor ≥ 4.0			Impact Factor ≥ 2.0 and < 4.0			Impact Factor < 2.0		
Journal	Impact factor	Sample %	Journal	Impact factor	Sample %	Journal	Impact factor	Sample %
† *Quarterly J. Ec.*	9.8	4.2	*Ecological Ec.*	3.9	1.5	*J. Banking & Fin.*	1.9	0.6
J. Fin. Ec.	5.9	1.9	*Energy Ec.*	3.4	0.1	*J. Int. Money & Fin.*	1.9	0.1
† *Econometrica*	5.8	2.5	*J. Health Ec.*	3.3	15.8	*European J. Political Ec.*	1.8	0.5
† *J. Political Ec.*	5.7	1.7	*Ec & Human Biology*	3.0	1.1	*J. Corporate Fin.*	1.8	0.4
† *American Ec. Rev.*	5.0	7.9	*J. Envir. Ec. & Manag.*	2.9	1.6	*European Ec. Rev.*	1.8	2.3
† *Rev. Ec. Studies*	4.7	1.5	*J. Urban Ec.*	2.9	3.9	*China Ec. Rev.*	1.8	1.1
J. Accounting & Ec.	4.7	0.1	*Food Policy*	2.8	0.0	*J. Ec. Psychology*	1.8	0.2
			J. Public Ec.	2.8	10.8	*J. Comparative Ec.*	1.7	1.3
			J. Development Ec.	2.8	10.3	*J. Ec. Behavior & Org.*	1.5	5.2
			J. Int. Ec.	2.7	1.1	*Ec. Education Rev.*	1.5	2.0
			World Development	2.7	5.8	*J. Empirical Fin.*	1.5	0.4
			J. Monetary Ec.	2.7	0.1	*Regional Sci. & Urban Ec.*	1.4	0.0
			J. Econometrics	2.3	0.4	*Labour Ec.*	1.4	8.3
			J. Fin. Stability	2.1	0.1	*Int. Rev. Ec. & Fin.*	1.4	0.5
			Resource & Energy Ec.	2.0	0.1	*Int. J. Industrial Org.*	1.4	0.6
						Information Ec. & Policy	1.1	0.1
						Explorations Ec. History	1.1	0.5
						Pacific-Basin Fin. J.	1.0	0.0
						J. Housing Ec.	1.0	0.0
						Ec. Modelling	0.9	0.4
						J. Japanese & Int. Ec.	0.8	0.1
						Ec. Letters	0.7	0.9
						Fin. Research Letters	0.7	0.0
						Int. Rev. Law & Ec.	0.5	0.4
						J. Applied Ec.	0.5	0.1

Note: Listing of journals constituting sample, divided into three categories by five-year impact factor. The first category of journals have a five-year impact factor above 4, the second from 2 to 4, and the third below 2. Five-year impact factors from *Thomson Reuters Journal Citation Reports 2015*. All listed journals contribute observations to sample, though some entries in the % of sample column round to 0.0%. Table omits listing of the 12 journals in 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%; together, they constitute less than 1.5% of the sample. Journals accessed via JSTOR site designated by †; these journals also happen to be regarded as the top-five general interest journals in economics; journals without this designation are published by Elsevier and accessed via the ScienceDirect website.

Appendix B: Supplementary Technical Details

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

B1. Bias Without Frequency Weighting

This section provides additional justification for our method of inverse frequency weighting (weighting observations by the inverse of the number of sniff tests in a table) used in our kernel-density and reduced-form-regression estimation. Letting $\bar{n}_\alpha = E_{t\in T_\alpha}(|t|)$, by the law of iterated expectations, the unweighted expectation can be written as

$$E_{i\in I_\alpha}(S_{i\alpha}) = E_{t\in T_\alpha}\left(\frac{|t|}{\bar{n}_\alpha}E(S_{i\alpha}\,|\,i\in t)\right). \tag{B1}$$

It is immediate that if tables are all the same size, then (1) equals (B1) because $|t| = \bar{n}_\alpha$.

To provide a more general comparison of the two expectations, we introduce the following reduced-form model. By equation (1), $S_{i\alpha}$ can be written

$$S_{i\alpha} = \pi_\alpha + u_{i\alpha}, \tag{B2}$$

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

$$E_{t\in T_\alpha}(E(u_{i\alpha}\,|\,i\in t)) = 0. \tag{B3}$$

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

$$E(\varepsilon_{i\alpha}\,|\,i\in t) = 0, \tag{B4}$$

whereas $\nu_{i\alpha}$ is unobserved table effect, which has zero mean across tables but can have nonzero mean within a table. Formally, letting $\eta_{t\alpha}$ denote expected value of the unobservable effect for table t, i.e.,

$$\eta_{t\alpha} = E(\nu_{i\alpha}\,|\,i\in t), \tag{B5}$$

we allow $\eta_{t\alpha}$ to be nonzero.

Substituting (B2)–(B5) into (B1) and rearranging yields

$$E_{i\in I_\alpha}(S_{i\alpha}) = \frac{1}{\bar{n}_\alpha}E_{t\in T_\alpha}(E(|t|\pi_\alpha + |t|\nu_{i\alpha} + |t|\varepsilon_{i\alpha}\,|\,i\in t)) \tag{B6}$$

$$= \pi_\alpha + \frac{1}{\bar{n}_\alpha}\mathrm{Cov}_{t\in T_\alpha}(|t|\eta_{t\alpha}). \tag{B7}$$

We see that the alternative expectation is biased relative to π_α, positively biased if $\eta_{t\alpha}$ covaries positively with table size $|t|$, negatively biased if $\eta_{t\alpha}$ covaries negatively with $|t|$.

B2. Sum of Frequency Weights

Summing the weights in equation (3) yields

$$\sum_{i\in I_\alpha} w_i = \frac{|I_\alpha|}{|T_\alpha|}\sum_{t\in T_\alpha}\left(\sum_{i\in t}\frac{1}{|t|}\right) = \frac{|I_\alpha|}{|T_\alpha|}\sum_{t\in T_\alpha} 1 = |I_\alpha|. \tag{B8}$$

B2. Extension of BUM Model to Multiple Removal Regions (BUMMR)

The density in (11) can be generalized to multiple regions with heterogeneous removal rates. To this end, partition the unit interval into K subintervals: $0 = \hat{\alpha}_0 < \hat{\alpha}_1 < \cdots < \hat{\alpha}_K = 1$. Let ρ_k be the removal rate in the subinterval $\alpha_k = [\hat{\alpha}_{k-1}, \hat{\alpha}_k)$. Applying Bayes Rule,

$$f(p_i;\mu,\omega,\vec{\rho}) = \frac{(1 - S_{i\alpha_k}\rho_k) f_1(p_i;\mu,\omega)}{1 - \sum_{k=1}^{K} (\rho_k - \rho_{k+1}) \left[\omega\hat{\alpha}_k + (1-\omega)\hat{\alpha}_k^{\frac{1}{1+m}} \right]}, \tag{B9}$$

where $\rho_{K+1} = 0$ by definition.

The denominator is the unconditional probability that a sniff test is note removed. This equals

$$\begin{aligned}
&\sum_{k=1}^{K} \Pr(p_i \in \alpha_k) \int_{\hat{\alpha}_{k-1}}^{\hat{\alpha}_k} (1-\rho_k) \frac{f_1(p_i;\mu,\omega)}{\Pr(p_i \in \alpha_k)} dp_i \\
&\quad = \sum_{k=1}^{K} (1-\rho_k) \int_{\hat{\alpha}_{k-1}}^{\hat{\alpha}_k} [\omega + (1-\omega) g(p_i;\mu)] dp_i && \text{(B10)} \\
&\quad = \sum_{k=1}^{K} (1-\rho_k) \left[\omega(\hat{\alpha}_k - \hat{\alpha}_{k-1}) + (1-\omega) \left(\hat{\alpha}_k^{\frac{1}{1+\mu}} - \hat{\alpha}_{k-1}^{\frac{1}{1+\mu}} \right) \right] && \text{(B11)} \\
&\quad = \omega\hat{\alpha}_K + (1-\omega)\hat{\alpha}_K^{\frac{1}{1+\mu}} - \omega \sum_{k=1}^{K} \rho_k \hat{\alpha}_k + \omega \sum_{k=1}^{K} \rho_k \hat{\alpha}_{k-1} \\
&\qquad - (1-\omega) \sum_{k=1}^{K} \rho_k \hat{\alpha}_k^{\frac{1}{1+\mu}} + (1-\omega) \sum_{k=1}^{K} \rho_k \hat{\alpha}_{k-1}^{\frac{1}{1+\mu}} && \text{(B12)} \\
&\quad = 1 - \omega \sum_{k=1}^{K} \rho_k \hat{\alpha}_k + \omega \sum_{k=0}^{K-1} \rho_{k+1} \hat{\alpha}_k \\
&\qquad - (1-\omega) \sum_{k=1}^{K} \rho_k \hat{\alpha}_k^{\frac{1}{1+\mu}} + (1-\omega) \sum_{k=0}^{K-1} \rho_{k+1} \hat{\alpha}_k^{\frac{1}{1+\mu}} && \text{(B13)} \\
&\quad = 1 - \omega \sum_{k=1}^{K} \rho_k \hat{\alpha}_k + \omega \sum_{k=1}^{K} \rho_{k+1} \hat{\alpha}_k \\
&\qquad - (1-\omega) \sum_{k=1}^{K} \rho_k \hat{\alpha}_k^{\frac{1}{1+\mu}} + (1-\omega) \sum_{k=1}^{K} \rho_{k+1} \hat{\alpha}_k^{\frac{1}{1+\mu}}. && \text{(B14)}
\end{aligned}$$

Equation (B10) follows from cancelling factors and substituting from (10), and (B11) follows from integrating. Equation (B12) follows from distributing 1 and cancelling a terms and from distributing ρ_k. Equation (B13) follows from $\hat{\alpha}_K = 1$ and from changing the indexing on two of the sums. Equation (B14) follows from $\hat{\alpha}_0 = 0$ and $\rho_{K+1} = 0$ by definition. Rearranging (B14) gives the denominator of (B9).

Generalizing the likelihood function from BUM(SR) in equations (12) and (13) to BUM(MR), we have

$$\ln L = \sum_{p_i \text{ exact}} \ln f(p_i;\mu,\omega,\vec{\rho}) + \sum_{p_i \in [\ell_i, u_i)} \ln \int_{\ell_i}^{u_i} f(p;\mu,\omega,\vec{\rho})\, dp \tag{B15}$$

$$
\begin{aligned}
&= \sum_{k=1}^{K} N_{p_i \in \alpha_k} \ln(1-\rho_k) - N \ln\left(1 - \sum_{k=1}^{K} (\rho_k - \rho_{k-1}) \left[\omega \hat{\alpha}_k + (1-\omega)\hat{\alpha}_k^{\frac{1}{1+\mu}}\right]\right) \\
&\quad + \sum_{p_i \text{ exact}} \ln\left(\omega + \frac{1-\omega}{1+\mu} p_i^{\frac{-\mu}{1+\mu}}\right) \\
&\quad + \sum_{p_i \in [\ell_i, u_i)} \ln\Bigg(\sum_{k=1}^{K} (1-\rho_k)\Big\{\omega\left[\min(u_i, \hat{\alpha}_k) - \max(\ell_i, \hat{\alpha}_{k-1})\right] \\
&\qquad\qquad + (1-\omega)\left[\min(u_i, \hat{\alpha}_k)^{\frac{1}{1+\mu}} - \max(\ell_i, \hat{\alpha}_{k-1})^{\frac{1}{1+\mu}}\right]\Big\}\Bigg).
\end{aligned}
\tag{B16}
$$