A meta-analytic investigation of p-hacking in e-commerce experimentation

Abstract

Randomized controlled trials—often called A/B tests in industrial settings—are an increasingly important element in the management many organizations. Such experiments are meant to bring the benefits of scientific rigor and statistical measurement to the domain of managerial decision making. But just as this practice is starting to reach widespread adoption, the problem of p-hacking—by which experimenters try several statistical analyses until they find one that produces a sufficiently small p-value—has emerged as a prevalent concern among statisticians, industrial practitioners, and the scientific community at large. In A/B testing in particular, experimenters have the ability to watch their data arrive in real time and stop experiments once their p-values reach a given threshold of statistical significance. Such behavior, which is known to inflate false discoveries, can cause managers to make costly mistakes with economically significant consequences. In this paper, we attempt to study the prevalence of this form of p-hacking in a sample of 2,482 experiments from 245 e-commerce firms conducted on a third-party A/B testing platform. After developing a statistical method to detect this effect, we apply it to our data and find (across several specifications) little to no evidence for p-hacking in our sample of experiments. We use counterfactual simulations to determine that if a modest effect of p-hacking were present in our dataset, our methodology would have high levels of power to detect it at our current sample size. In addition to outlining a robust method for detecting p-hacking in similar datasets, our finding serves as a valuable data point in an increasingly important discussion on how economic agents use data and statistics for strategic decision making.

1 Introduction & Background

1.1 Experimentation and A/B testing. In the first half of the 20th century, an industrial researcher named Ronald Fisher published several books that would go on to be some of the most influential works in all of statistical science. By advocating for the use of p-values and significance testing, these works popularized the use of experiments for instrumental purposes. Over time, Fisher's statistical methods—in particular his emphasis on randomization—have come to revolutionize how stakeholders across a number of disciplines make consequential decisions (Bellemare and Bloem, 2017, Bothwell and Podolsky, 2016, BuiltWith, 2019, Hamermesh, 2013, Kleven, 2018, Kohavi, 2019, Sanders and Halpern, 2014).

As the practice of data driven decision making and analytics gained prominence over the last several decades (Brynjolfsson et al., 2011, Brynjolfsson and McElheran, 2016, McAfee et al., 2012), the use of Fisherian experiments for strategic decisions have become much more common. This is particularly true in *digital* business, where the marginal cost of both service innovation and delivery is relatively low (Mitchell et al., 2003). Dozens of SaaS platforms have launched in recent years that enable firms to run experiments on their websites and apps, sometimes at no cost[1] (Gartner, 2019). A growing body of research on digital experimentation—now widely known as "A/B testing" among software, marketing, and e-commerce companies—has emerged in recent years, including work from academics in many fields such as economics, marketing, information systems, and statistics (Azevedo et al., 2018, Heck et al., 2019, Kohavi et al., 2013, Liu and Chamberlain, 2018, Mislavsky et al., 2019, Peysakhovich and Eckles, 2018). In industrial contexts, A/B testing has emerged as a key component of the corporate innovation process, allowing firms to statistically compare competing strategies about product-market fit, pricing, messaging, targeting, and user experience. Several studies have documented the use of A/B testing for these strategic purposes among large enterprises, online merchants, and technology startups (Kohavi and Longbotham, 2017, Koning et al., 2019, Miller and Hosanagar, 2018). While the largest technology companies have used online experiments for years, the use of A/B testing is growing dramatically among firms of all types. This can be seen in the results of a 2018 survey of more than 200 companies with at least $500 million in annual revenue, in which 74% of respondents indicated they either already use or plan to use A/B testing in the near future (Virzi, 2018).

[1] Among others, "Optimize", a service on the Google Marketing Platform, currently offers a completely free A/B testing service for small websites.

1.2 p-hacking & false discovery. While experimental methods have continued to proliferate, there has been a renewed interest in the downsides of common statistical procedures for analyzing experimental results. Much of this research focuses on the shortcomings of null hypothesis significance testing (NHST) and its emphasis on p-values, which has been the dominant paradigm for conducting statistical tests in academic science for decades (Huberty, 1993).[2] It has long been known that statistical research based on significance testing is likely filled with many (if not a majority of) false positive results due, in part, to a selection effect induced by the peer-review and publication process (Franco et al., 2016, Ioannidis, 2005, Rosenthal, 1979). However, in addition to this passive selection effect present in the *reporting* of experimental results, more recent studies have highlighted how flexibility in the *design and analysis* of experimental results can dramatically inflate empirical false discovery rates (Gelman and Loken, 2013). This phenomenon, by which researchers change their data sampling procedures or analytic techniques to obtain "statistically significant" results, has come to be known as "p-hacking" (Simmons et al., 2011). Concern about the prevalence of both publication bias and p-hacking in many areas of academia—including psychology, economics, and biostatistics—has led some scholars to characterize the current state of scientific inquiry as being in the midst of an epistemological "crisis" (Dougherty, 2008, Dreber and Johannesson, 2019, Earp and Trafimow, 2015, Gelman and Loken, 2016).

Despite this crisis, NHST and p-values have come to predominate much of the statistical software used throughout the A/B testing industry. A particularly pernicious form of p-hacking enabled by digital experimentation is the practice of continuous monitoring, whereby experimenters regularly check a test's p-value and only end an experiment when a sufficiently small ("statistically significant") value is obtained. Many testing platforms will explicitly highlight when an experiment's p-value dips below the conventional significance level of 5%, with some going so far as to notify their users when this threshold is met. Many practitioners are aware of the pitfalls of continuous monitoring in online experiments, encouraging the use of sample size calculators or Bayesian methods in place of classical NHST (Borden, 2014, Draper, 2016, Miller, 2010). At larger organizations, a significant amount of resources has been allocated to the problem of developing valid statistical methods that are robust to continuous monitoring. While research on related problems dates back to at least Wald (1945), methods for dealing with continuous monitoring in A/B testing continues to be an active area of industrial research. Microsoft, Walmart,

[2] See Schneider (2015) for a discussion on the origin of the modern, commonly-used practice of NHST and its relationship to the foundational statistical methods developed by Fisher, Neyman, and Pearson.

Twitter, AirBnb, Uber, and Optimizely have all published recent work on the topic (Abhishek and Mannor, 2017, Deng et al., 2016, Feng, 2017, Lu, 2016, Overgoor, 2014, Pekelis et al., 2015).

However, the rise of low-cost experimentation platforms means many firms using A/B testing software might not have the statistical sophistication necessary to combat problems of multiple comparisons and, instead, must rely on interpreting the statistics reported by testing platforms.[3] Given that academic researchers, often with doctoral degrees and graduate training in statistics, have been known to engage in p-hacking behavior (Hartgerink, 2017, Head et al., 2015, Leggett et al., 2013, Masicampo and Lalande, 2012, Perneger and Combescure, 2017), it is reasonable to ask whether analysts in corporate environments, using similar experimental techniques (though perhaps with different incentives) make similar methodological errors. Answering this question can have far-reaching implications for both how researchers and managers understand the value A/B testing and statistical methodologies in corporate environments. To our knowledge, only one recent working paper has addressed the subject of p-hacking in similar contexts. Berman et al. (2018) study a large sample of firms with data from more than 2,000 experiments on a major commercial A/B testing platform. They claim more than 70% of experimenters in their sample stop their experiments when the p-values drop below the 10% significance threshold. While several aspects of the context in that paper differ from ours (discussed in the context of our own results in more detail later), it demonstrates that p-hacking is a costly and potentially widespread behavior on commonly-used testing platforms. If p-hacking behavior is indeed a widespread phenomenon, many firms would be justified in reevaluating the way they use experiments for business decisions.

1.3 Behavioral explanations and moral hazard. As we hope to add to our understanding of p-hacking, particularly in corporate environments, we find it is useful to consider the possible behavioral antecedents of p-hacking behavior.[4] Given that notions of statistical significance, p-values, and hypothesis testing are not intuitive for many students of statistics (Aquilonius and Brenner, 2015, Greenland et al., 2016, Hubbard, 2011), some practitioners seem to misunderstand the goal of experimental analysis as achieving statistical significance rather than recovering true model parameters (Szucs, 2016, Ware and Munafò, 2015). This can cause analysts to engage in a number of "questionable research practices" without full knowledge of how such practices undermine their own conclusions (John et al., 2012, Sijtsma, 2016). Analytic flexibility combined

[3] It is known that investments and IT software require complementary investments in organizational capacities, which are not always coincident (Brynjolfsson and Hitt, 1996).

[4] See Smith et al. (2019) for a discussion on the primary causes of the "replication crisis" in social science more broadly.

with this misunderstanding that induces a bias toward statistical significance can easily be seen to result in p-hacking behavior (Carp, 2012).

An alternative, and perhaps more interesting, explanation for p-hacking behavior is the presence of misaligned incentives. It is easy to see how in academic science, the imperative for researchers to publish in prestigious journals, combined with reluctance of many journal editors to publish null results, can result in strong incentives to engage in p-hacking behavior (Krawczyk, 2015, Woolf, 1986).[5] And while public interest might be better served if academics minimized their incidence of false discoveries, individual researchers often face little risk by publishing "significant" results that later turn out to be flawed. The same potential for moral hazard in experimental analysis is also present in the context of A/B testing at private companies. If employees are rewarded for the number of "significant" findings they produce, but verification of these findings is inherently costly for managers, principle-agent dynamics provide one rationalization for p-hacking behavior among economic agents (Baker, 1992, Holmstrom and Milgrom, 1991). On the other hand, one could argue that the negative consequences of p-hacking are larger in industrial settings than in academia, perhaps mitigating the effects of moral hazard. As discussed earlier, A/B tests are frequently used for purposes instrumental to a firm's economic strategies, informing managerial decisions about product variations, service delivery, user experience, and marketing campaigns. The results of experiments in these domains—particularly among internet-scale companies—can often have significant effects on profitability (Hern, 2014). To the extent that one believes employee and firm incentives are aligned, we would expect an industrial practitioner to face at least some economic incentive to avoid engaging in p-hacking behavior.[6] As such, it appears not obvious that the same problems which plague academic research would also be found in industrial contexts.

This discussion motivates the current project, in which we set out to investigate the empirical prevalence of p-hacking among economically-motivated decision makers. We proceed by conducting a meta-analysis of p-values from a large sample of A/B tests conducted by e-commerce merchants. Our analysis will exploit the fact that, if experimenters do regularly stop their tests right

[5] Such behavior need not be intentional for the effects of these incentives to play a role, as it is possible for researchers to make many *post-hoc* justifiable design choices that result a slight bias toward specifications with significant results. A more formal analysis of this phenomenon in econometric research is discussed in Spiess (2018), who casts the misalignment of social and private incentives in a principle-agent framework.

[6] Research by Aral et al. (2012) suggests that firms who adopt analytics and IT practices are more likely to adopt more efficient human capital management practices; this provides some reason to believe that firms who use A/B tests—which is likely associated with adoption of IT and analytics practices in general—might be better at aligning organizational incentives.

when their p-values reach 0.05, we would expect to see a jump—or discontinuity—in the number of p-values observed right below this threshold. As such, we have developed a robust statistical methodology for analyzing this type of discontinuity in p-value distributions. We apply this technique to the distribution of p-values from 2,482 experiments conducted by 245 firms and find little to no evidence for the incidence of p-hacking in our dataset. While it is not strictly possible to prove a null hypothesis, this does not mean our findings are uninformative about p-hacking behavior among e-commerce companies (Abadie, 2018). In addition to demonstrating the robustness of our null result across many specifications, we also use counterfactual simulations to demonstrate that, if a modest effect of p-hacking did exist, our statistical methodology would have the power to detect it at our current sample size. Despite finding a null result, this research makes a valuable contribution by presenting both a theoretically-motivated method for detecting discontinuities in p-value distributions and empirical evidence about how real-world firms use and deploy statistical tools for managerial decision making.

2 Empirical Context

To study the problem of p-hacking behavior among e-commerce firms, we have gathered data from a third-party A/B testing platform (subsequently referred to as "the platform"). We provide more detailed summary statistics in the following section, but at a high level we have access to a population of two-armed experiments (i.e., literal "A/B" tests) conducted on the platform between 2014 and 2017. But before proceeding, several aspects of the platform interface are important to discuss. We will outline the process by which a firm creates an experiment, the data that are visible to firms during ongoing experiments, and the process by which firms stop an experiment.

2.1 Testing Interface. Much like any other third-party testing services provider, firms using the platform's A/B testing technology must create an account and configure their website's codebase before an experiment can be run.[7] Once their website is configured, firms can use the platform's WYSIWYG interface or custom CSS/Javascript injection to create an intervention to be tested in an experiment. For example, a firm may want to change the image on their homepage or measure the effect of advertising a promotion on checkout pages. Once an experiment is created and deployed, the platform automatically manages all session randomization, analytics measurement, statistical calculations, and reporting of results.

[7]This almost always requires (1) ensuring a callback function is executed when important website actions are performed by site visitors (e.g., product views, conversions, etc.); and (2) the installation of a Javascript tag that allows the testing provider to manipulate website elements and measure session characteristics.

In terms of what the platform measures, by default, eight dependent variables are analyzed for every experiment. These are: conversion rate, session revenue, new visitor conversion rate, add-to-cart rate, cart abandonment rate, page views, session duration, and bounce rate. Prior to starting an experiment, firms are allowed to choose one or more "target metrics", which correspond to the outcome(s) being targeted by the intervention. If no target metric is explicitly selected by the firm, the platform sets the target to "conversion rate". In 89% of experiments in our sample, conversion rate is specified as a target metric; in 82%, it is the *only* target metric. The most other common target metrics are session revenue and add-to-cart rate.

2.2 What data do firms have available during an experiment?. The primary interface by which firms view the results of their experiments is an online dashboard listing the outcome metrics, their associated effect sizes (known as "lift" in digital marketing), standard errors, and "confidence" levels (defined as 1 minus the p-value). A stylized version of the dashboard which is functionally similar to the real interface is shown in Figure 1. Experimenters can click on a specific metric to see the time-series history of the effect size over time, but by default they merely see an up-to-date snapshot overview like the one shown in the example interface.

Figure 1: Stylized test result dashboard interface

Actionable		**Confidence**	**Lift**	**Standard Error**
Time on site		96%	0.30%	±0.12
Pending				
Revenue		15%	-$5.30	±7.89
Conversion rate	✔ Target Metric	38%	-0.02%	±0.02
New visitor conversion rate		41%	0.00%	±0.01
Add to cart		61%	+0.21%	±0.15
Pageviews		37%	+0.89	±2.93
Abandonment		83%	+1.02%	±0.74
Bounce rate		92%	-6.32%	±3.52

One characteristic about the interface worth noting is that there is always a badge indicating which "target metric" the experimenter specified at the beginning of the test. This reinforces to the firm the metric that they pre-specified as their primary performance indicator for the experiment. Also, important for our research question, as soon as a variable crosses above 95% confidence (i.e., its p-value dips below 0.05), it is moved from the "pending" portion of the dashboard to the

Table 1: Data Summary

Firm-Level	Mean	Median	Std. Dev.	Min	Max
Experiments per firm	10.1	4	14.8	1	95
Experiment-Level					
Length of experiment (days)	44.6	28.1	45.0	0.1	399.3
Number of sessions	201,842.857	37,770.5	429,062.358	60	6,034,939
Conversion rate	0.098	0.043	0.161	0.0	1.0
Terminal p-value	0.427	0.41	0.311	0.0	1.0

"actionable" portion and shown in a different color. Through the use of these design cues, the platform clearly reinforces the importance of the 5% significance threshold and, in a not entirely subtle way, encourages firms to react to their experiments when they reach this threshold.

We begin our analysis by analyzing firm response to the p-value on the *conversion rate* metric. In addition to being the most commonly specified "target metric" in our sample, conversations with employees at the platform indicate conversion rate has an outsized importance among e-commerce websites generally and is often the ultimate objective of a marketing intervention. This is also reinforced by the platform's setting of conversion rate as the default target metric and the fact that very few experimenters select alternative metrics.

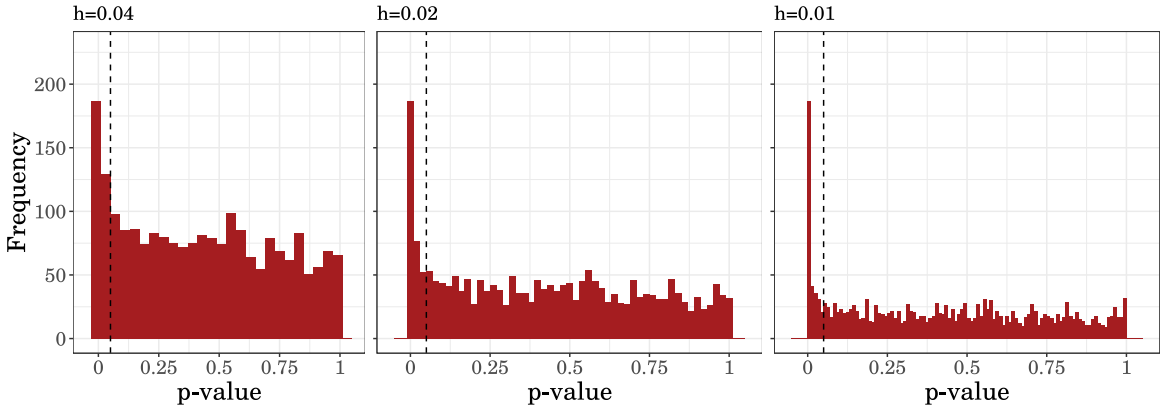
3 Data & Preliminary Analysis

In this section, we describe more quantitative details of the experiments in our dataset and start to examine the empirical distribution of their associated p-values. As mentioned, all experiments occurred between 2014 and 2017 with data collection transpiring between 2017 and 2018. We have summarized some essential information about the data in Table 1. In total, we have 2,482 experiments in our sample from 245 firms. As can be seen from the distribution of the number of sessions in each experiment, there is considerable range in our sample. While the median experiment in our data has more than 37,000 customer visits in its lifetime, the largest experiment had more than 6 million. This is indicative of the type of firm that is a typical customer of the platform: a mid to large sized e-commerce company with thousands of visitors per day to their website.

3.1 Empirical distribution of p-values. One way to observe p-hacking behavior in our dataset is to look at the empirical distribution of p-values. If we assume that, for some proportion of experiments in our sample, firms were continuously monitoring their primary p-values (i.e., those associated with each test's *conversion rate*) and stopping experiments when they dropped below 0.05, we should expect to see a disproportionately large amount of tests with p-values below this threshold than above it. We can begin looking for evidence of this artifact by examining histograms of the raw distribution of p-values. In Figure 2, we have plotted histograms of our data for three different bin widths (denoted h), along with a dashed vertical line at 0.05 to facilitate inspection at this critical threshold.

A fundamental challenge in estimating properties of empirical density functions—which we will have to address later—is their sensitivity to the choice of modeling parameters. As can be seen in these graphs, different characteristics of the density function are visible at different resolutions. While it appears there are significantly more p-values just below 0.05 in the plot with bandwidth $h = 0.04$ (far left panel), this does not appear to be true if we look at the graph with $h = 0.02$ (middle panel).

Figure 2: Density of Terminal p-values



Notes: Histogram plots of raw data shown for various bandwidths h. Dotted line is place at 0.05 to highlight the anticipated p-hacking threshold.

3.2 Preliminary analysis: Poisson regression. As a preliminary attempt to investigate the nature of our data near the 0.05 threshold, we fit a Poisson regression discontinuity model to the bin counts of our histograms. While this method has some drawbacks, it is useful to begin our analysis here for several reasons. First, because we are ultimately analyzing count data (the *number* of p-values in a given region of the distribution space), Poisson regression is a natural, well-

motivated, first-pass approach for modeling our outcome of interest (Efron, 2012). Second, this method provides a useful visualization of the statistical uncertainty associated with the histogram heights near the 0.05 threshold. Lastly, this exercise is useful for illustrating the difficulties associated with empirical density estimation in general and highlights some of the peculiarities of working with p-value distributions in particular.

There are many modeling choices that we must specify before we can test for discontinuity at 0.05 using Poisson regression. To begin, we consider a histogram of our data with fixed bin width of $h = 0.005$. The centers of each histogram bin will serve as our independent variable x, and the heights of each bin will serve as our dependent variable y. In this regression, we use a quadratic basis for x to allow for curvature in the density estimate. Additionally, because we want to allow for discontinuous behavior above and below $x = 0.05$, we also include an indicator vector $I\{x < 0.05\}$ among our independent variables. Lastly, because we are most keenly interested in the behavior of this function near $x = 0.05$, we restrict our analysis to only consider data in the range $x \in [0.005, 0.095]$. Formally, we express this specification as a generalized linear model with log link function to be estimated using the principle of maximum likelihood (McCullagh and Nelder, 1989):

$$\log\left(\mathbf{E}[y \mid x]\right) = \beta_0 + \beta_1 x + \beta_2 x^2 + \gamma I\{x < 0.05\} \tag{1}$$

We show the results of this model in Table 2. We have also plotted the predicted mean and 95% confidence intervals of the Poisson regression on top of the underlying histogram data in Figure 3. As can be seen visually in the graphic, as well as statistically by inspecting the coefficient on our indicator variable ($\gamma = -0.18, p = 0.439$), this model is unable to detect a discontinuity in our data at the 0.05 threshold.

While we believe there may be good reasons for believing the results of this particular model, it is important to consider the many "researcher degrees of freedom" that went into determining the test statistic (Simmons et al., 2011). Even if Poisson regression is an appropriate model specification for our research question, there are many modeling choices that have material impact on the fit of this model and the resulting statistical outcomes. In Table 3, we identify at least six different degrees of freedom that any researcher using regression discontinuity methods for density discontinuity analysis must specify before statistical results can be observed.[8] Rather than to belabor justifying the particular modeling choices we have made above, we argue that a more

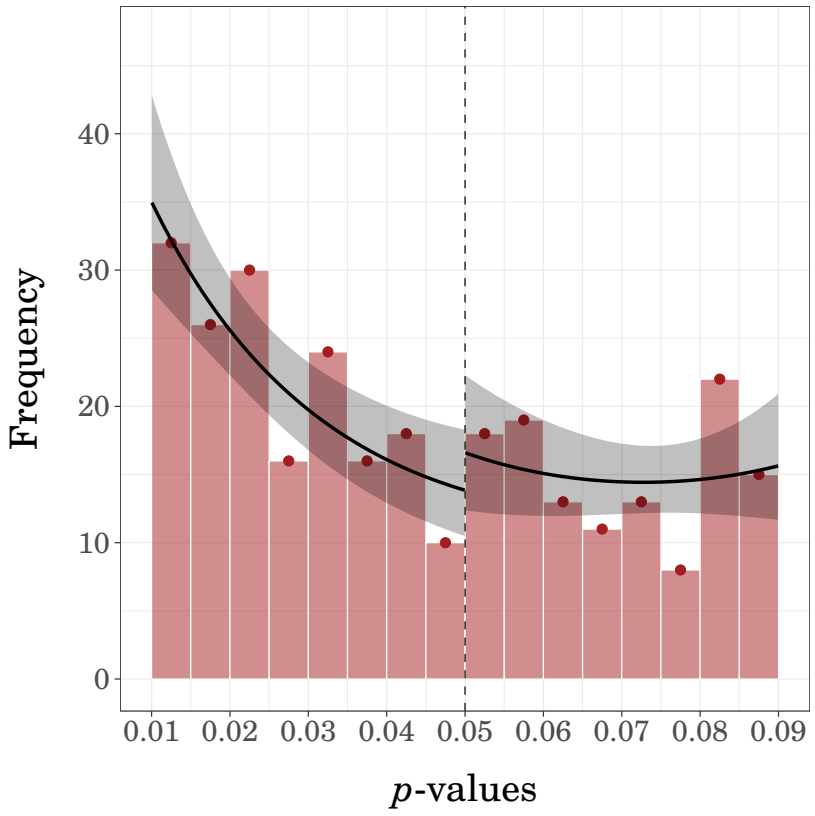
[8]These are in addition to other universal degrees of freedom, like those that factor into a researcher's data collection and inclusion procedures.

Table 2: Poisson regression parameter estimates

	Dependent variable
	log (bin height)
β_0 (Intercept)	4.10^{***}
	(0.33)
β_1	-39.44^{***}
	(9.03)
β_2	271.04^{**}
	(88.72)
γ	-0.18
	(0.23)

Notes: $p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Table displays maximum likelihood based estimates (and associated standard errors) of the parameters specified in the linear model shown in Eq. 1. The primary coefficient of interest is γ, which measures the estimated difference between the empirical density of p-values on either side of the 0.05 threshold.

Figure 3: Poisson regression fit



Notes: Visualization of Poisson regression estimates of bin height (black line) with 95% confidence bands shown (grey regions). The empirical histogram (red bars) used to generate the bin height data (red dots) is shown in background, with bin widths set to $h = 0.005$.

principled approach would be to choose a model with fewer degrees of freedom to begin with.

As a starting place, one might consider models that come with data-adaptive bandwidth se-

Table 3: Researcher degrees of freedom when using regression discontinuity for density discontinuity

Free parameter	**Notes**
The upper and lower limits of the data range to include	When combined with parametric or semi-parametric estimators, the behavior of a distribution near the endpoints of its data range can dramatically affect inference near the threshold of interest.
The decision to model our frequency data using a standard histogram plot, represented by a single point at the center of each bin	Standard histograms are simply one of many possible non-parametric density estimators for obtaining frequency distributions from an empirical sample of data. See Bourel and Ghattas (2012) for a comparison of many such methods.
The parameter h for determining the histogram's bin widths	Different values of h can significantly change the shape of a density estimate
The decision to use a polynomial basis for smoothing the histogram data	As opposed to other reasonable specifications that allow for more flexible smoothing, e.g., locally weighted polynomials (Cleveland and Devlin, 1988), or generalized additive models (Hastie, 2017).
The order of the polynomial in this basis	Gelman and Imbens (2019) demonstrate how flexibility in this parameter can be particularly pernicious for controlling false discovery in regression discontinuity analyses.
The specification of the discontinuity term	That is, whether to include a single additive indicator variable $I\{x < 0.05\}$ or to interact this indicator with the polynomial terms, allowing for independent fits on either side of the threshold.

lection procedures (Otsu et al., 2013). However, in addition to reducing the degrees of freedom in our model specification, there are at least two reasons for developing a tailor-made statistical approach for studying discontinuities in p-value distributions in particular. First, because p-value distributions exhibit extremely skew (perhaps asymptotic) behavior as $x \to 0^+$, existing approaches may not handle the peculiarities of our empirical context well. Since the threshold of interest in our research question, 0.05, is relatively close to this margin, even typically robust techniques can be sensitive to how close to zero we allow our sample to be. Second, most off-the-shelf discontinuity detection methods use non-parametric, local density estimation techniques. While there is nothing wrong with these approaches in general, in our context we can incorporate prior knowledge about the shape of p-value distributions, giving us a *global* model for the density of our data. Local density estimation methods ignore this global information, which can lead to smaller effective sample sizes.[9] If we can propose a well-specified model for what our distribution of p-values should be in the *absence* of p-hacking, we can use this approach to better quantify

[9]In more words, global density models consider all of the data in a sample simultaneously, without ignoring any data or down-weighting data far from the threshold of interest. This can be useful if a researcher has reason to believe in the fidelity of their model restrictions. Because we are working with p-values, we can draw upon statistical theory for specifying a well-motivated null model for our data.

how the behavior of our data near 0.05 deviates from this null model. In the following section, we propose such a model and use it to develop a statistical test for p-hacking behavior in our sample.

4 Density Discontinuity Analysis

We have attempted to design a statistical procedure to test precisely our phenomenon of interest (discontinuity in the density of p-values near 0.05) in an easily-interpretable, transparent, and robust way that makes use of all the data in our sample. This procedure has two arbitrary input variables: the number of mixture components K and a bandwidth parameter h. We discuss robustness to each of these modeling choices below.

While there are several components to our test, its basic outline is not unlike that of any statistical test based on classical significance testing. We outline the five main steps of the procedure below:

1. Estimate a well-specified, parametric model of the empirical p-value density function, $\hat{f}$, assuming no p-hacking in the data generating process.[10]
2. Specify a test statistic, S, that takes as input a density of p-values, f, and attempts to measure the presence of a discontinuity near 0.05.
3. Derive a "null distribution" of what the statistic S would look like if $\hat{f}$ were the true model: $\hat{S} = S \mid f \sim \hat{f}$
4. Compare the empirical values of the test statistic to the null distribution to obtain a p-value for the test of discontinuity. If we wish to use formal hypothesis testing procedures, we can ask if the empirical value falls in the extreme portion of the null sampling distribution (i.e., above the 95th percentile for $\alpha = 0.05$), reject null model and conclude evidence for a discontinuity in the density of p-values exists.
5. For robustness, repeat steps 3 and 4 by varying the choice of researcher-selected parameters (in this case, the density bandwidth parameter h).

We describe each of these steps in detail below.

4.1 Modeling the empirical distribution of p-values. If we assume our data were generated through valid (i.e., non-p-hacked) experimental procedures, we can draw upon existing statistical theory about how a "typical" distribution of p-values should look. For an experiment

[10] Throughout this section, we use hat notation (e.g., $\hat{f}$) to denote *estimated* quantities; variables without hat decorations refer to theoretical quantities.

where the null hypothesis is true, i.e., there is no difference in conversion rates between treatment arms, theory predicts the p-value to be uniformly distributed on the unit interval. However, when we look at the results of an experiment, we do not know *ex ante* whether the null hypothesis is true or if, on the other hand, a non-zero effect size is present. If we look at the results of many experiments in a meta-analysis of p-values, some p-values will correspond to tests where the null hypothesis is true (and therefore be uniformly distributed), but for another portion the alternative hypothesis will be true (when the effect size is non-zero). In these cases, two-sided p-values will tend to cluster near zero (since such results have a lower probability of occurring when assuming the null hypothesis). This explains the rough shape of the p-value distributions plotted in the previous section.

This discussion motivates the modeling of the distribution of p-values as a hierarchical mixture coming from one null component (with uniform distribution to model the null effects) and another component to model the results of non-null effects. In line with literature in both statistics and biostatistics on the modeling of meta-analytic distributions of p-values (Allison et al., 2002, Gronau et al., 2017, Nettleton et al., 2006, Parker and Rothenberg, 1988, Pounds and Morris, 2003, Tang et al., 2007)[11], we model the non-null portion of our density as a mixture of beta distributions, with the parameter restrictions $\alpha < 1$ and $\beta > 1$.[12]

This model can be formally described using the following data generating process. Let x_i denote the value of a single p-value observation drawn i.i.d. from some underlying distribution with density $f(x_i \mid \theta)$. This densiy f depends on a parameter vector θ, which is comprised of a set of sub-variables $\theta = (\pi, a, b)$ that characterize the components of the following finite mixture model with $K+1$ components.[13] In the first stage, a mixture component $k_i \in \{0, 1, \dots, K\}$ is drawn from a categorical (or multinoulli) distribution with parameter vector $\pi = (\pi_0, \dots, \pi_K)$:

$$k_i \mid \pi \sim \text{Categorical}(\pi_0, \dots, \pi_K) \ ,$$

subject to a natural ordering and unitary sum constraint:

$$\pi_0 > \pi_1 > \cdots > \pi_K \ \text{ and } \ \sum_k \pi_k = 1 \ .$$

Conditional on knowing an observation's mixture component k_i, x_i will be distributed either uniformly (for $k_i = 0$) or as a Beta random variable with component-specific parameters (a_{k_i}, b_{k_i})

[11] This literature is primarily concerned with modeling the results of genome-wide association studies, in which the effects of thousands of genes need to be simultaneously estimated through meta-analytic techniques. See Cai and Sun (2017) for a general discussion of these techniques.

[12] These inequalities merely constrain the distribution space to only those functions that could plausibly model a sample of p-values with positive skew.

[13] See McLachlan et al. (2019) for a detailed overview of finite mixture models.

(for $k_i > 0$):

$$x_i \mid k_i = 0 \sim \text{Unif}(0, 1)$$

$$x_i \mid k_i > 0, a, b \sim \text{Beta}(a_{k_i}, b_{k_i})$$

Given a set of observations $x = (x_1, \ldots, x_N)$, one can estimate the parameter vector θ that best fits this data using the principle of maximum likelihood.[14] To proceed, we first integrate over our uncertainty in the mixture component k_i to obtain a marginal probability, f, which will be our base model for our data's underlying density function:

$$f(x_i \mid \theta) = \int_k f(x_i \mid \theta, k_i = k) dF(k) = \pi_0 + \sum_{k=1}^{K} \pi_k f_k(x_i) \tag{2}$$

In this notation, f_k is the beta density function corresponding to the k-th component:

$$f_k(x) = \frac{1}{B(a_k, b_k)} x^{a_k} (1 - x)^{b_k - 1}$$

where B is the beta function.

The likelihood function for our model can then be calculated by taking the product of the likelihood over each data point in the empirical vector x:

$$\mathscr{L}(\theta \mid x) = \prod_{i=1}^{N} f(x_i \mid \theta) = \prod_{i=1}^{N} \left(\pi_0 + \sum_{k=1}^{K} \pi_k f_k(x_i) \right)$$

allowing us to derive the data's log-likelihood as:

$$\begin{aligned} \ell(\theta \mid x) &= \log \mathscr{L}(\theta \mid x) \\ &= \log \left(\prod_{i=1}^{N} \left[\pi_0 + \sum_{k=1}^{K} \pi_k f_k(x_i) \right] \right) \\ &= \sum_{i=1}^{N} \log \left(\pi_0 + \sum_{k=1}^{K} \pi_k f_k(x_i) \right) \end{aligned}$$

The maximum likelihood parameter estimates can then be obtained by maximizing this function: $\hat{\theta}_{MLE} = \arg\max_\theta \ell(\theta \mid x)$.

While we have, until this point, described a generic model for an arbitrary number of components K, we must now fix this parameter in order to estimate the mixture model. For the primary results reported here, we use a model with $K = 2$ beta components (three in total, counting the uniform component). However, we note that model fit and all derived results are not materially affected by using larger choices for this parameter.[15]

[14] One can also esimate this model using a fully Bayesian approach, in which distributions over the parameters $\theta = (\pi, a, b)$ must be specified *a priori*; see Tang et al. (2007) for such an example.

[15] By imposing an ordering constraint on the mass of each mixture components, so that $\pi_0 > \pi_1 > \cdots > \pi_K$, the effect of this parameter on model outcomes diminishes as K grows. This is unlike the order of the polynomial used in regression designs (like that discussed in Section 3), which can substantially affect the shape and flexibility of a model specification in unpredictable ways as this parameter varies.

Having fixed the functional form of our model, we can now fit its parameters to our dataset; these estimates are reported in Table 4. The model described above lends itself quite well to graphical visualization; the MLE fit of our model on top of our the empirical histogram has been plotted in Figure 4. In grey, we have generated a histogram of empirically observed p-values from our dataset; on top of this, we show the best-fit maximum likelihood estimate, as specified above. This graph is broken up into a null component (f_0, in blue) and a composite "alternative" component ($f_a = f_1 + f_2$, in red), corresponding to data generated from a mixture of positively skewed beta densities. Since the density function outlined here is clearly continuous at 0.05, it provides a theoretically-motivated null model[16] to facilitate hypothesis testing in the following sections.

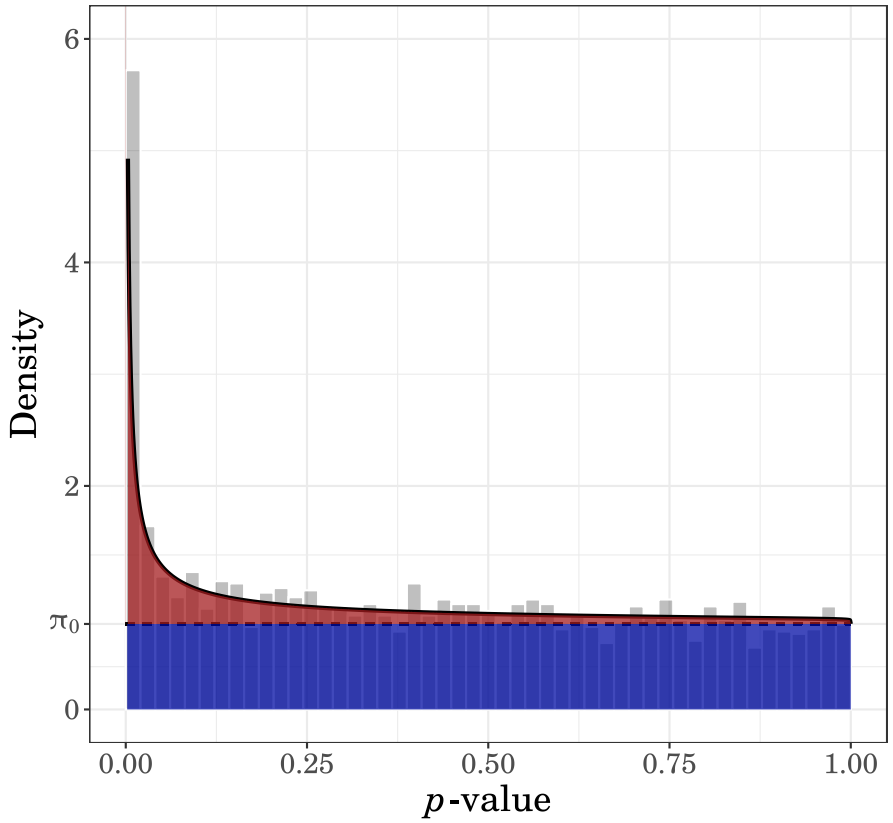
Table 4: MLE Parameter Estimates

Parameter	Estimate
π_0	0.771
π_1	0.199
π_2	0.029
a_1	0.277
a_2	0.014
b_1	1.62
b_2	6.44

4.2 A simple statistic for detecting discontinuity. To test for a discontinuity at the 0.05 threshold in our data, we will use a simple, common-sense approach: counting the number of p-values in our sample that are close to either side of 0.05. However, this approach requires that we introduce a bandwidth parameter, h, which determines exactly what "close" means in this context. While we could use a data-adaptive approach (so that, generally, larger sample sizes allow for smaller bandwidths), we will instead fix the bandwidth ahead of time. The reason for this is that data-adaptive approaches are prone to small-sample biases that can have unpredictable effects on statistical tests dependent on this parameter. In Appendix C, we demonstrate how the statistical procedure we use to detect discontinuity with fixed bandwidth exhibits higher power and more regularity than the same procedure with adaptive bandwidth. We are able to show how the small-sample effects of a data-adaptive approach are still prominent at our given sample size of $N = 2{,}482$. As such, we believe there is good reason to select a fixed bandwidth that would allow us to detect our primary effect of interest. Researchers studying meta-analytic p-value distributions in academic research have previously pointed to a prevalence of p-values in the 0.04-0.05 range as

[16] "Null" in the sense of there being *no* discontinuity

Figure 4: Beta-uniform mixture MLE fit and empirical histogram



Notes: Histogram of empirically observed data (shown in **grey**) with bin width $h = 0.02$. Under our model's assumptions, p-values can be thought of as coming from a null (uniform) component with probability $\hat{\pi}_0 = 0.771$ (visualized in **blue**); alternatively, p-values can arise from an *alternative* component (with probability $1 - \hat{\pi}_0 = 0.229$ (visualized in **red**), which is a mixture of beta distributions.

indicative of p-hacking behavior (de Winter and Dodou, 2015, Simonsohn et al., 2014).

As such, we will proceed by assuming a fixed the bandwidth parameter of $h = 0.01$. We can attempt to minimize the arbitrariness of this choice by also reporting results of our analysis for alternative specifications in which this parameter equal to half ($h = 0.005$) and twice ($h = 0.02$) this value. This method of reporting robustness to bandwidth is commonly used in regression discontinuity designs, as first suggested by McCrary (2008).

Conditional on a given value of h and a threshold of interest τ (0.05 in this case), we motivate the definition of a test statistic comparing the number of p-values on either side of this threshold as follows. We are formally interested in testing the continuity of the density function:

$$\lim_{h \to 0+} f(\tau - h) = \lim_{h \to 0+} f(\tau + h)$$

For a finite (which is to say, any) sample, it is impossible to evaluate the infinitesimal limits above. To get around this, we sum (or integrate) the value of the density function near either

side of the threshold $\tau = 0.05$, allowing us to ask whether the difference between the number of observations above and below this threshold is larger than would be expected by random chance. With a fixed value of the bandwidth parameter h, we can straightforwardly compare the number of observations that fall in the range $[\tau - h, \tau]$—denoted N_l (i.e., *left* of the threshold)—to the number of observations in the range $[\tau, \tau + h]$—denoted N_r (*right* of the threshold). We formalize the difference between these values as a test statistic with the following formula:

$$S(f) = N_l - N_r \tag{3}$$

4.3 Null distribution of test statistic. To derive the null distribution of the test statistic above, first note that due to the slope of the density function near zero, it is not appropriate to merely assume an equal number of observations on either side of the 0.05 threshold. Even in non-p-hacked data we would expect to see slightly higher number of observations below 0.05 than above it. The question we need to answer is whether there is a *disproportionate* difference in our empirical sample. We can account for the shape of the density near the 0.05 threshold by calculating what *proportion* of the distribution is on either side of the threshold within the $[\tau - h, \tau + h]$ window. In particular, let π_l be the fraction of the distribution in this window on the left side of the critical threshold, and let π_r be the proportion on the right side:

$$\pi_l = \int_{\tau-h}^{\tau} f(x)dx, \quad \pi_r = \int_{\tau}^{\tau+h} f(x)dx$$

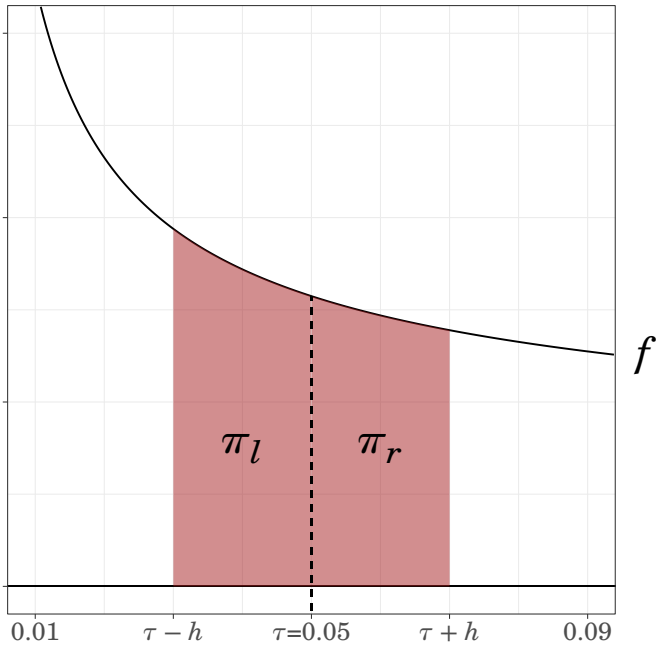
A visualization of these parameters is provided in Figure 5.

One can use these probabilities to model the null distribution of *counts* above and below the threshold as two binomially distributed random variables, drawn from the entire sample of size N:

$$N_l \sim \text{Binomial}(N, \pi_l), \quad N_r \sim \text{Binomial}(N, \pi_r)$$

The mass function for test statistic specified in Eq. (3) can then be derived by summing up all ways in which these two binomials can differ by some value $k \in \mathbb{N}$; because k can be positive or

Figure 5: Visual representation of test statistic parameters



Notes: The regions marked π_l and π_r (visualized in **red**) represent the relative proportions of the p-value distribution immediately below and above the threshold of interest, $\tau = 0.05$. Along with a given sample size N and specified bandwidth parameter h, these parameters fully identify the null distribution of our test statistic.

negative, these cases must be modeled separately:

$$
\begin{aligned}
P[S = k] &= \begin{cases} \sum_{i=0}^{N} P[N_l = i + k]\ P[N_r = i], & \text{if } k \geq 0 \\ \sum_{i=0}^{N} P[N_l = i]\ P[N_r = i + k], & \text{otherwise} \end{cases} \\
&= \begin{cases} \sum_{i=0}^{N} \binom{N}{i+k} (\pi_l)^{i+k} (1-\pi_l)^{N-(i+k)} \binom{N}{i} (\pi_r)^{i} (1-\pi_r)^{N-i}, & \text{if } k \geq 0 \\ \sum_{i=0}^{N} \binom{N}{i} (\pi_l)^{i} (1-\pi_l)^{N-i} \binom{N}{i+k} (\pi_r)^{i+k} (1-\pi_r)^{N-(i+k)}, & \text{otherwise} \end{cases}
\end{aligned}
\tag{4}
$$

In our case, where we have parametrically estimated f, the parameters of this distribution—π_r, π_l—can be explicitly estimated by plugging in the functional form of f in Eq. 2, along with its estimated parameters $\hat{\theta}$ from Table 4. The MLE estimate for the π_r parameter, for example, is

given by:

$$\begin{aligned}\hat{\pi}_r &= \int_{\tau}^{\tau+h} \hat{f}(x)dx \\ &= \int_{\tau}^{\tau+h} \left[\hat{\pi}_0 + \sum_{k=1}^{K} \hat{\pi}_k \hat{f}_k(x)\right] dx \\ &= \pi_0 \int_{\tau}^{\tau+h} dx + \sum_{k=1}^{K} \pi_k \int_{\tau}^{\tau+h} \hat{f}_k(x)dx \\ &= \hat{\pi}_0 h + \sum_{k=1}^{K} \hat{\pi}_k \left[I(\tau+h;\hat{\alpha}_k,\hat{\beta}_k) - I(\tau;\hat{\alpha}_k,\hat{\beta}_k)\right]\end{aligned} \tag{5}$$

where $I(x;\alpha,\beta)$ is the regularized incomplete beta function, the CDF of the beta distribution; π_l can be similarly derived.

We are now in a position to formally state the primary hypothesis test of interest for determining whether there is a discontinuity in the distribution of p-values in our sample near the 0.05 threshold (thereby indicating the incidence of p-hacking among firms in our sample). Because p-hacking behavior of the type described earlier is only consistent with observing an excess mass of p-values *below* the 0.05 threshold, we employ a one-sided hypothesis test; specifically, we want to test whether our statistic of interest (defined by subtracting the number of p-values below the threshold from those above the threshold) is strictly positive: $S > 0$.

To define a formal hypothesis test for this particular case, we assume under the null hypothesis that our data are drawn from the continuous distribution specified in Eq. (2) and that our test statistic is less than or equal to 0. The condition that $S > 0$—indicating a disproportionate number of p-values below 0.05—serves as our alternative hypothesis:

$$\begin{cases} H_0: \quad S \leq 0 \iff \lim_{h\to 0+} f(\tau-h) \leq \lim_{h\to 0+} f(\tau+h);\ S \sim \hat{S}, f \sim \hat{f} \\ H_a: \quad S > 0 \iff \lim_{h\to 0+} f(\tau-h) > \lim_{h\to 0+} f(\tau+h) \end{cases} \tag{6}$$

4.4 Definition of p-value for test statistic. We can derive a frequentist p-value from our test statistic by assuming the null hypothesis H_0, i.e., that the continuous distribution model (parameterized by maximum likelihood estimates) is true. This value will represent the probability that we would observe a test statistic (i.e., a difference between the number of test results just above and below the 0.05 threshold) as large or larger than the observed statistic, assuming our distribution follows the null beta-uniform mixture model described above. If S^* is the empirically observed difference between the number of observations above and below τ, and $\hat{S}$ is the random variate of our test statistic with the null distribution under the MLE model, the formula for this

p-value can be expressed as:

$$p\text{-value} = P\left[S^* \leq S \mid H_0 : S \sim \hat{S}\right] = \int_{S^*}^{\infty} dP(\hat{S}) = \sum_{k=S^*}^{\infty} P[\hat{S} = k] \tag{7}$$

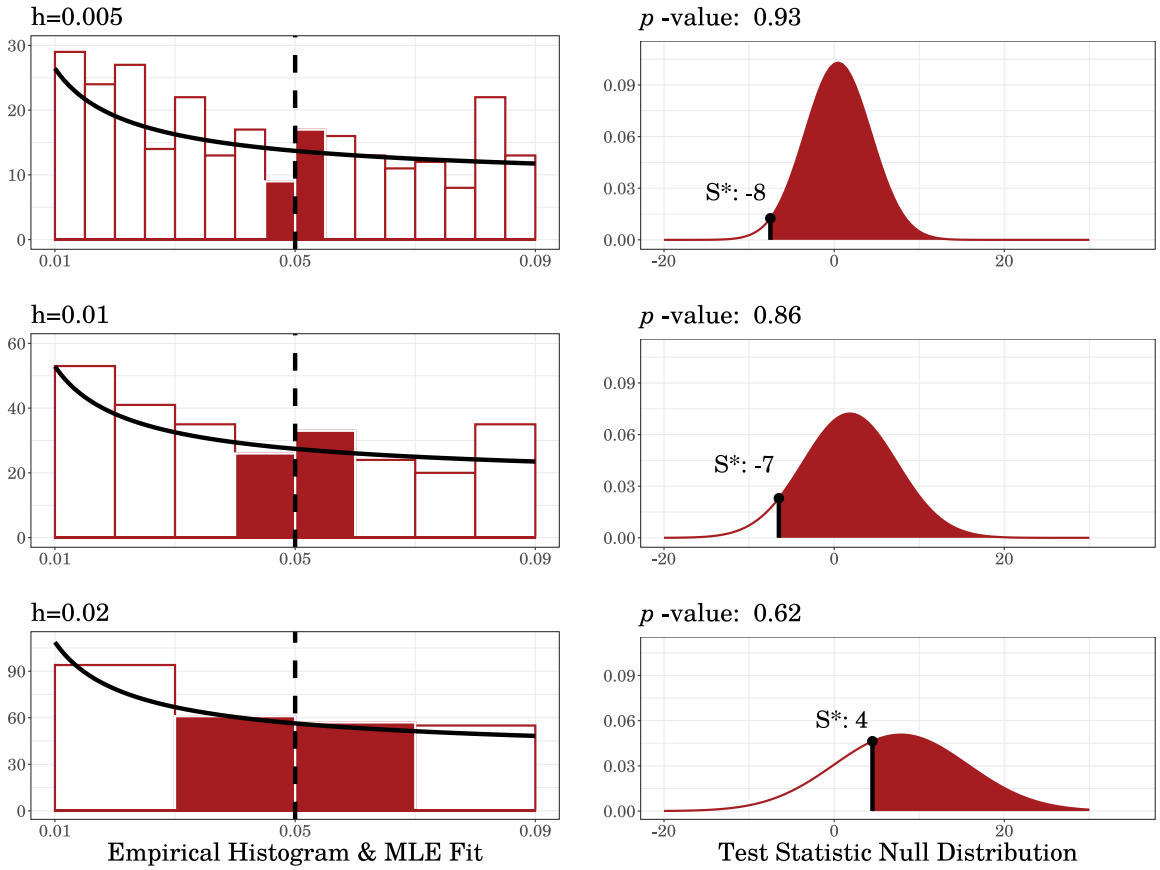
4.5 Empirical Results: Compare observed difference to null distribution. To summarize our entire testing procedure once more: Conditional on a known bandwidth value h, we define our test statistic as the difference between the number of observations within h units of either size of our critical threshold $\tau = 0.05$. We then derived the null distribution of this statistic under the assumption that our data were generated by the parametric beta-uniform mixture model. Having estimated this parametric model by MLE in Section 4.1, we can now compare the empirically observed test statistic with its theoretical distribution under the null and derive a corresponding p-value. As mentioned earlier, we consider our primary test to be that when $h = 0.01$, but for robustness we also report results for $h = 0.005$ and 0.02.

These results are displayed in Figure 6. Each row of the figure corresponds to a separate calculation for each of the bandwidth choices $h = 0.005, 0.01, 0.02$. The left panel shows histograms of the empirical distribution with binwidth set to h (the corresponding h value is shown above each histogram). The primary characteristic of these histograms being considered is he difference between the bin counts immediately above and below the 0.05 threshold; these have been colored in with solid red to highlight the portion of the data that is used for each test. Additionally, the MLE fit of the beta-uniform mixture model shown as a solid black line above the histograms. The right side of the panel shows the null distribution of the test statistic (based on the MLE fit), along with the empirically observed test statistic S^*. This value corresponds directly to the difference in the highlighted histogram bin heights above and below 0.05. The p-values in this figure are calculated using the formula derived earlier in equation (7), and are displayed above the graphs of the test statistic distributions.

As can be seen, the p-values from our tests are all well above any reasonable significance level for any value of the bandwidth parameter ($p = 0.93, 0.86, 0.62$). In statistical terms, we are unable to reject the null hypothesis of continuity in the underlying density function of p-values near the 0.05 threshold. In behavioral terms, these results provide no evidence for the form of p-hacking described earlier on the part of the firms in our sample.[17]

[17] If we had performed two-sided hypothesis tests (rather than the one-sided test described in Eq. (6)), the p-values for the smaller bandwidths are closer to (but not below) conventional levels of statistical significance—indicating that, if anything, there may be a disproportionate number of experiments with p-values just *above* 0.05. But evidence for this conclusion would be very weak, as it seems to depend critically on the choice of bandwidth h. Interpreting both these results together, it is reasonable to conclude there is little to no evidence for *any* type of discontinuity at the 0.05

Figure 6: Main Results



Notes: Our primary statistical test measures the difference between the number of p-values above/below the 0.05 threshold and compares this value to a theoretically derived null distribution of this statistic, based on fitting a beta-uniform mixture model to the observed data. We perform this test three times, once for each of the bandwidth values $h = 0.005, 0.01, 0.02$. Empirically observed histograms for each of these bandwidth values is shown in the left column (white/red bars), along with the MLE-fit beta-uniform mixture model specified in Eq. 2 (black line). In the right column, for each h, we plot (a smoothed version of) the corresponding null distribution of our test statistic (derived in Eq. 4) and the empirically observed value S^* (defined in Eq. 3). Taking the proportion of the null distribution that lies above S^* gives us the p-value for our discontinuity test (see Eq. 7 in text). No p-value among the outcomes we observe ($p = 0.93, 0.68, 0.62$) approaches conventional levels of significance, indicating we cannot reject a null hypothesis that assumes continuity in the underlying density function of p-values.

5 Power Analysis

It is important to recognize that absence of evidence is not evidence of absence. As in all frequentist hypothesis testing, it is not possible for us to "test for" a null hypothesis. In this case, we cannot say for sure that the experiments in our dataset were not p-hacked, merely that the method we developed was unable to reject the null hypothesis at conventional significance levels. What we can do is contextualize the results we have observed in our dataset through power analysis.

Because we are investigating a unique behavioral phenomenon, we do not have the benefit of simple statistical theory to estimate the power of our experiment. In traditional power analyses

threshold, and exceedingly weak evidence for the specific type of discontinuity associated with p-hacking behavior.

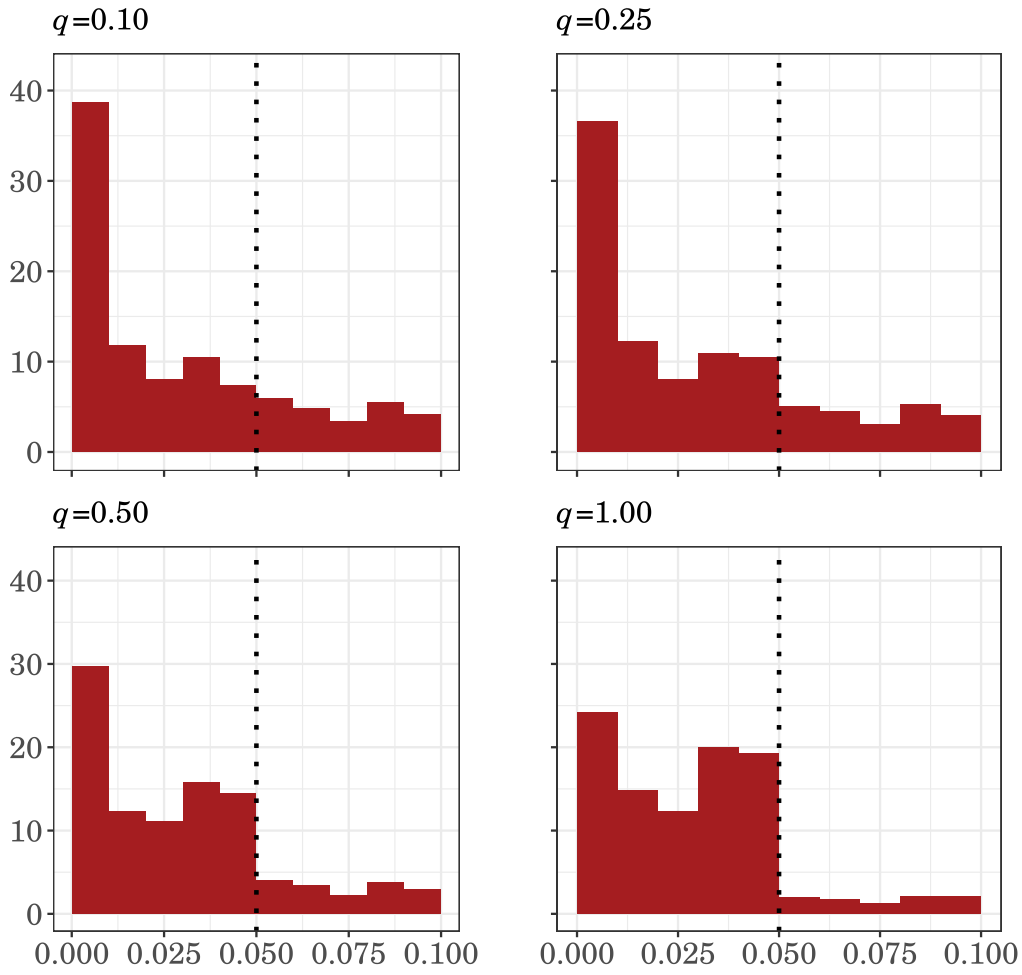
(e.g., the difference between the means of two populations), the "effect size" is well defined with known theoretical distribution under the null hypothesis. However, in our context, the effect we are trying to detect—the presence of p-hacking behavior—does not have an obvious analytical relationship to the test statistic we derived in Section 4. As such, we will have to develop our own method of evaluating the power of this test. We propose using counterfactual simulations to estimate power. Before proceeding, we will need (1) a rigorous definition of the effect we are trying to detect and how to quantify its "effect size", and (2) a method for observing what our data would look like if such an effect were truly present. We discuss each of these below.

Effect size of p-hacking. We argue the most meaningful way to think about "effect size" in our context is as *the proportion of experiments that were intended to be p-hacked*. The word "intended" is necessary because not every experiment will have a p-value that dips below 0.05 within its lifespan. In our own data, for which we have calculated a time series of p-values at 24 hour increments over the course of each experiment, only 29.9% of experiments have a p-value that dips below 0.05 at any point in its time series. For more than 70% of the experiments in our sample, even if the firms monitored the p-value every day and intended to p-hack the data by stopping early, they would not see a p-value below 0.05 in the natural lifespan of the experiment. Note, however, that firms would not know *ex ante* whether a given experiment will have a p-value that dips below 0.05 or not. Thus, we believe the most natural way to quantify the "amount" of p-hacking in a sample such as ours is the percentage of experiments for which an experimenter *would* stop early if given the chance. We will refer to this quantity as the "effect size" or "extent of p-hacking", represented by $q \in [0, 1]$.

Counterfactual simulation of p-hacking behavior. We will only be able to assess the power of our statistical test if we are able to observe or model its results when applied to a set of experiments that have truly been p-hacked (i.e., for which $q > 0$). In this section, we outline a method for simulating p-hacking behavior in our data. As mentioned above, we are able to observe the contemporaneous p-value for each of the experiments in our dataset at 24 hour intervals (i.e., the p-value each firm would have seen if they looked at the testing dashboard at that time in the experiment). This gives us the ability to "counterfactually p-hack" our data by walking through each experiment day-by-day and "stopping" it if we see a p-value below 0.05 ("stopping" in this case just means ignoring any data collected after that point). To simulate the effect of p-hacking on the distribution of p-values in our set of experiments, we can apply this approach to each experiment in the sample and collect the simulated terminal p-values. In Figure 7, we have illustrated

what the resulting p-value distributions looks like when applying this p-hacking procedure for different effect sizes in our dataset: $q = 0.10, 0.50, 1.00$. In each case, we randomly selected a proportion q of experiments from our entire sample, and then simulated the effect of p-hacking as described above. For experiments that were not p-hacked, their original terminal p-values remain unchanged. As can be clearly be seen, the phenomenon hypothesized earlier, in which p-hacking leads to a discontinuity in the distribution of p-values, is clearly visible for larger effect sizes. To assess the statistical power of our methodology, we must now ask how often our test rejects the null hypothesis in this simulated data and how small q must be before the test is unable to reliably detect the p-hacking behavior.

Figure 7: Histograms of p-values for counterfactually p-hacked experiments



Notes: Histograms are generated by randomly selecting a proportion q of experiments in our dataset to p-hack. For each experiment to be p-hacked, we replaced its empirically observed terminal p-value with the value that would have been observed if the experimenter checked p-values once every 24 hours and terminated the test on the first day its p-value dropped below 0.05. A vertical dotted line is placed at the 0.05 threshold.

Before proceeding, we make two comments about the simulation procedure described above. First, in the real world, the notion of sample size flexibility can entail both (1) stopping an experiment early and (2) gathering *more* data based on early results (Simmons et al., 2011). We call this form of p-hacking, in which analysts collects more data than they otherwise would have, *data extension*. Note our simulation procedure makes no allowances for data extension. To simulate p-hacking, we merely take the number of days in an experiment that were actually recorded, and never simulate the collection of additional data if a p-value is not significant at its observed terminal value. As such, when we describe our effect size q as the proportion of experiments p-hacked, a procedure that allows for both early stopping and data extension would actually result in even larger effects than we have simulated. This implies we are *underestimating* the ability of our test to detect the presence of p-hacking.

Second, the procedure as described simulates the behavior of a firm that opens their testing dashboard once a day (every day, at the same time), checks the p-value of their experiment, and stops the experiment if the value is below 0.05. The *frequency* at which a p-value is checked will clearly alter the results of our analysis. Thus, it should be noted that an effect size of q in our context means the proportion of experiments that were intended to be p-hacked *by monitoring p-values once a day*. Real p-hacking behavior would obviously be less deterministic than the behavior we are simulating. However, we argue that this method does capture many of the essential dynamics of real p-hacking behavior, while keeping the problem tractable. Note that if we believe real firms monitor their experiments *more frequently* than once a day, then our simulations should underestimate the consequences of this behavior on the distribution of terminal p-values.[18]

Results. Having defined our effect size and described a method for observing the downstream consequences of a non-zero effect in our data, we are now in a position to formally assess the power of our discontinuity test. In particular, we would like to be able to answer the question: for a given sample size N (i.e., number of experiments) and a given non-zero effect size $q \in (0, 1]$, what is the probability that the statistical method outlined in Section 4 will detect this effect?[19] In Appendix B, we outline an algorithm for using bootstrap aggregation to average over the sampling variation associated with our simulation procedure. This method allows us to approximate the reliability of our statistical test for detecting p-hacking behavior for a given sample size, N, and effect size,

[18] We are thus dramatically understating our ability to detect truly *continuous* monitoring, whereby an experimenter observes their results after every observation.

[19] Since our method is based on the paradigm of null hypothesis significance testing, we can only answer this question for a given false positive rate, α. As is convention in much of the scientific literature, we set $\alpha = 0.05$.

q.

To get a full picture of how well our discontinuity detection method works across a range of assumptions about its input data, we have calculated its power on a grid of sample sizes $N \in \{100, 250, 500, 1000, 2500, 5000, 10000\}$ and effect sizes $q \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1\}$[20]. The results are shown in a contour plot in Figure 8a, which shows the power of our testing procedure at each combination of N and q. We first remark how the shape of the contours match with what would be expected of a good statistical test: namely, that power increases when either sample size or effect size increases. Second, we have also plotted a dashed line at $N = 2842$ to highlight the power of our test as applied to our current sample of experiments. While the power to detect a small effect size of $q = 0.10$ is only around 20%, at $q = 0.20$ the power of our test jumps up to 76% (very near the conventional 80% power level often used in statistics). For any effect size above approximately $q = 0.30$, our analysis suggests our test has power of 95% or higher.

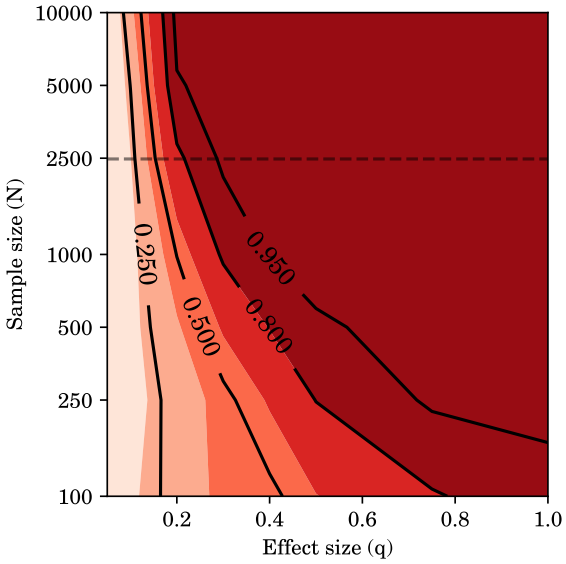
The results of these calculations give us greater confidence in the null results we found earlier. For modest effect sizes, in which only 20%-30% of experiments are (intended to be) p-hacked, the test we developed appears to detect these effects with high levels of reliability. Note that for effect sizes much smaller than 20%, since this quantity merely represents the proportion of experiments *intended* to be p-hacked, only a very small number of p-values (near 2%-3%) are actually affected by this behavior in a typical simulation. (We can see in Figure 8a that the power of our test at this level of effect grows quite slowly as sample size increases up to 5,000 or even 10,000 experiments.) Thus, while we cannot conclusively rule out the presence of p-hacking in our sample, we can say with reasonable confidence that the effect would have to be relatively small if it were present.

6 Alternative Specifications

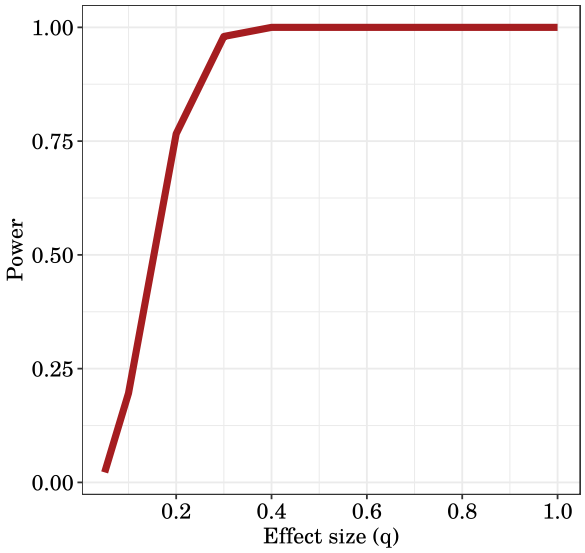
6.1 Off-the-shelf methods. Up to this point, we have looked for a discontinuity in our data using a naive polynomial regression method (Section 3) and a technique developed specifically for our research context (Section 4). We have argued for the utility of our method, but also acknowledge that the results of an entirely novel technique may be hard to interpret in isolation. As an additional robustness check to our main results, it can be helpful to see if our findings remain the same if we use existing, off-the-shelf techniques for discontinuity detection. In this section, we use the "simple local polynomial estimation" technique developed by Cattaneo et al. (2018a), which is a robust, non-parameteric, boundary-adaptive method for analyzing empirical density functions.

[20]The results shown here correspond to the statistical test with fixed bandwidth $h = 0.01$; the power contours for the tests with $h = 0.005$ and $h = 0.02$ are shown in the Appendix, Figure 11

Figure 8: Power (sensitivity) analysis of method for detecting p-hacking



(a) Power (z-axis) of our primary statistical test at varying sample sizes N (y-axis) and varying effect sizes q (x-axis). (Dotted line shown at sample size of our dataset, $N = 2,482$.)



(b) Power (y-axis) of our primary statistical test, broken out for our given sample size of $N = 2,482$ and varying effect sizes q (x-axis).

As with all density estimation techniques, this test requires the specification of a bandwidth value h. We use the MSE-optimal bandwidth suggested by Cattaneo et al. (2018b) to derive a statistic T for the test of discontinuity in our data at the 0.05 threshold. With this T-statistic we derive both 2-sided (testing for *any* discontinuity at the 0.05 threshold) and 1-sided p-values (testing specifically for the expected form of the discontinuity, that there are more experiments with p-values below 0.05 than above 0.05). We report the results of these calculations in Table 5. This procedure yields p-values $p = 0.466$ and $p = 0.767$ for the two-sided and one-sided tests, respectively. As such, we see no evidence for a discontinuity at 0.05 in our data.

At this point, we have looked for a discontinuity in our data using three distinct techniques, including a naive Poisson regression, our tailor-made method, and the non-parametric technique used in the section. The conclusions of all our tests so far have been qualitatively consistent, with none of them finding evidence of a discontinuity in our p-value distribution near 0.05.

6.2 Alternative Outcome Metrics. Until now, we have analyzed the distribution of p-values in our sample corresponding to the statistical test measuring the treatment effect of each experiment on *conversion rates*. While conversion rate is the most commonly specified target metric in our sample, when firms are viewing the results of their experiments on the testing platform's

Table 5: Simple local polynomial test for discontinuity in density of p-values at 0.05

Optimal bandwidths $(\hat{h}_l, \hat{h}_r)$	$(0.011, 0.011)$
Robust T-statistic	0.7294
2-sided p-value $H_a : f(c_-) \neq f(c_+)$ $P(\lvert T\rvert > t)$	0.466
1-sided p-value $H_a : f(c_-) > f(c_+)$ $P(T < t)$	0.767

Notes: We implement this test using the recommended second-order polynomial fit ($p = 2$), a triangular kernel, and jackknife estimator for variance estimation. Other choices of these parameters yield qualitatively similar results.

dashboard, they can see the statistics corresponding to all eight dependent variables measured, with only a small visual distinction between goal and non-goal metrics. As shown in the snapshot of the example dashboard in Figure 1, the "time on site" treatment effect is statistically significant but the goal metric of "conversion rate" is not. If we assume firms react equally to each of the eight outcome metrics on their testing dashboards—and hypothesize they stop their experiments if any single p-value dips below 0.05—we should see evidence for this behavior in the empirical distribution of p-values. But in this case, rather than looking for a discontinuity in the distribution of p-values from a single outcome metric, the operant p-value for a given experiment (i.e., the p-value most strongly influencing stopping behavior) will be the *smallest* of all eight p-values visible to a firm at a given point in time. By examining the distribution of smallest terminal p-values observed for each of the experiments in our sample, we can determine if there is a disproportionate amount of these p-values below the 0.05 threshold.

To be precise, let X_m be the terminal p-value corresponding to the m-th outcome metric from a given experiment. We will now consider the distribution of "minimal p-values", which is just the smallest p-value observed across all eight outcomes at the end of an experiment, which we denote by $X_{(1)} = \min\{X_1, \ldots, X_8\}$. To detect a discontinuity in the empirically observed distribution of this variable, we will use the same method as before: we first assume that that our data are drawn from a continuous distribution, derive the maximum likelihood fit of this distribution to our data, use this distribution to derive a null distribution of the test statistic, and determine whether the empirically observed test statistic falls in the extreme regions of this null distribution.

There is one key difference between this analysis and that from Section 4, which is that our model for the density of minimal p-values must be adapted to account for the application of the minimum operator in our definition of $X_{(1)}$. To do so, we would like to use existing theory about order statistics to derive a closed-form expression for the distribution of $X_{(1)}$. However, we must first make the simplifying assumption, that the p-values for each of the 8 outcome metrics are independent and identically distributed (as a beta-uniform mixture random variable). Strictly speaking, it is not likely to be true that the outcomes of different metrics from the same experiment would be completely independent. As a simple example, a treatment's effect on "time on site" is likely to be correlated with its effect on "pageviews". That being said, modeling this dependence (or any variation between the shapes of the p-value distributions) should not be necessary for the purposes of conducting our statistical test for discontinuity. So long as we are able to derive a continuous distribution that we have good reason to believe will fit the distribution of "smallest p-values", the main purpose of this exercise is to model the shape of the null distribution near 0.05. We demonstrate below how, despite this simplification, the resulting parametric distribution does fit our data reasonably well.

Given the condition that X_m are i.i.d., $X_{(1)}$ will be distributed as the smallest order statistic of eight draws from a p-value distribution, each of which can be modeled as before using the beta-uniform mixture (see Eq. 2). In Appendix A, we derive a closed-form expression for the density of the smallest order statistic of an arbitrary number (M) of samples from the beta-uniform mixture model with an arbitrary number (K) of components. This function can be expressed as:

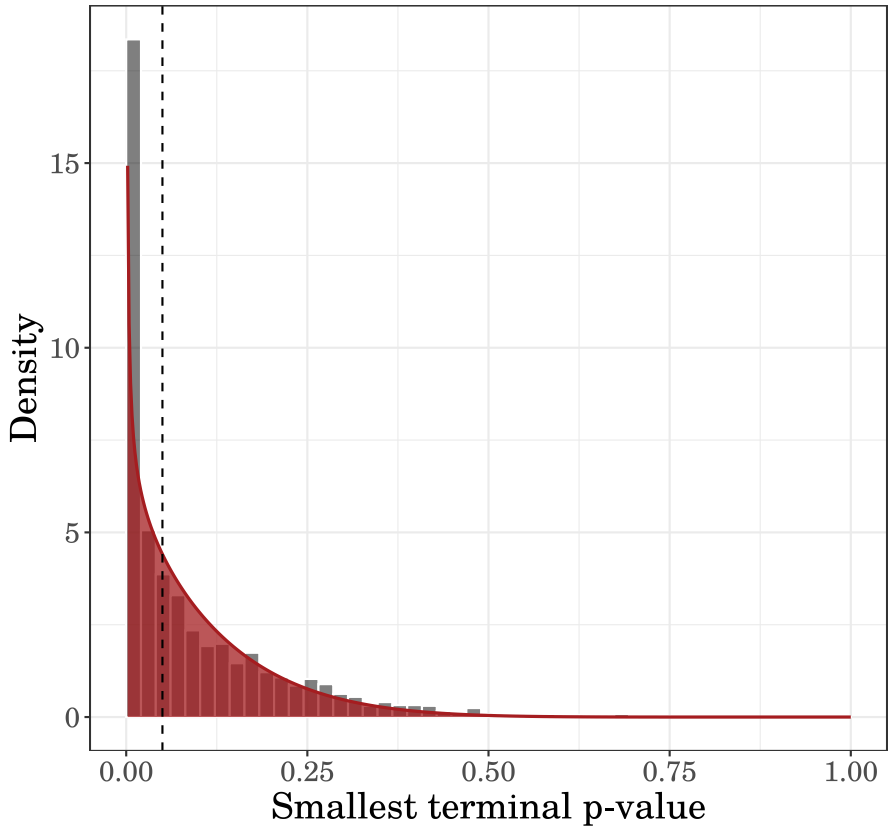
$$f_{(1)}^{M}(x) = \sum_{j_0+j_1+\cdots+j_K=M} (-1)^{1+M-j_0} \binom{M}{j_0, j_1, \cdots, j_K} \left(\prod_{k=1}^{K} (\pi_k F_k(x))^{j_k} \right) \left(\sum_{k=1}^{K} \frac{j_k f_k(x)}{F_k(x)} \right) \quad (8)$$

where $\pi = (\pi_1, \pi_2, \pi_3)$ represent the mixture probabilities and each beta-distributed component f_k is depends on two parameters $\{\alpha_k, \beta_k\}$.

For the purposes of modeling our data, we use two beta components (as in Section 4) and will be modeling the smallest order statistic from a sample of eight independent draws from this mixture distribution. As before, we use maximum likelihood to fit the parameters of this distribution. In Figure 9, we have plotted the MLE fit of this function on top of the empirical histogram of minimal p-values from the experiments in our sample.

Having calculated the MLE fit of the distribution of minimal p-values, we can perform the same set of hypothesis tests for discontinuity that we performed in Section 4. Namely, we calculate the

Figure 9: Distribution of minimial p-values, across all eight outcome metrics

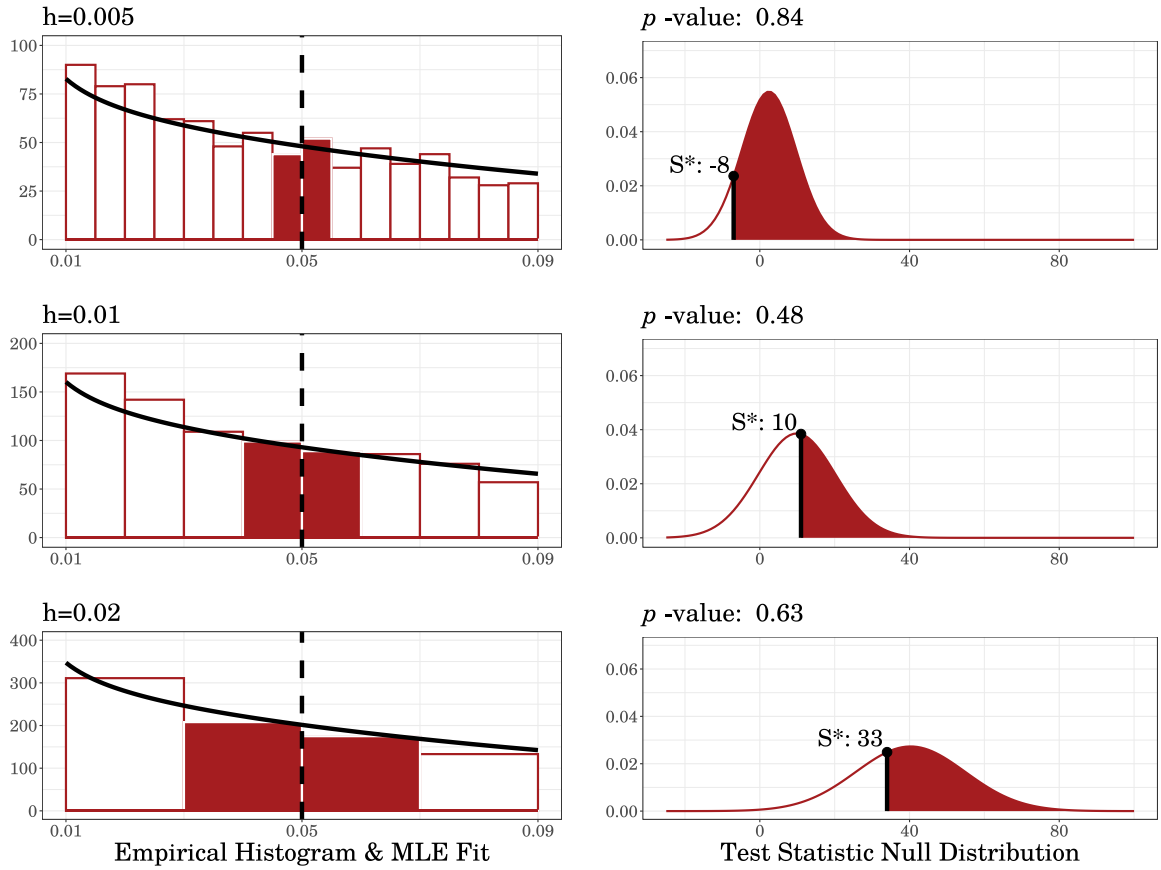


Notes: Histogram of empirically observed data is shown in **grey**. These data are generated by considering, for each A/B test in our sample, only the smallest terminal p-value across the eight outcome metrics at the end of the experiment. Assuming our data were not p-hacked, we derived an analytical equation for this distribution (see Eq. 8) and estimated its parameters using maximum likelihood; the shape of this fit is visualized in **red**, and serves as the null model used for hypothesis testing (see Figure 10).

p-value from Eq. 7 using the calculations derived in Eqs. 4 and 5. These results are displayed in the same fashion as before, for varying values of $h = 0.005, 0.01, 0.02$, in Figure 10. By both visually inspecting the histograms and observing the large p-values associated with our test for discontinuity ($p = 0.84, 0.48, 0.63$), we again find no evidence for p-hacking under this specification. These findings—which make use of all eight p-values visible to firms throughout their experiments—appear to reaffirm the conclusion from Section 4, which found little to no evidence of p-hacking by firms in response to conversion rate statistics.

6.3 Different data inclusion criteria. In the analyses discussed above, we have treated each experiment in our sample as an independent observation. While this simplification may make sense for most of our experiments, there is a proportion for which we can conclude that independence is likely violated. There are a number of experiments in our sample which, though

Figure 10: Robustness test for discontinuity, among meta-distribution of all outcomes



Notes: This figure contains a similar set of tests to those in Section 4, except rather than considering only the distribution of p-values corresponding to *conversion rates*, we consider the distribution of p-values derived by taking the smallest observed p-value for each of the eight outcome metrics for which the testing platform calculates statistics. The null model (shown as black lines in the left column) is described in more detail in Figure 9.

registered as separate tests on the testing platform, appear to start and end at nearly the same time. Given that the stopping time for each test in a group of coterminous experiments is clearly correlated, treating each of the tests as an independent observation may inflate our effective sample size.[21] There are a number of ways of dealing with this independence violation, but to provide the most conservative demonstration of our null results, we will attempt to develop a specification that is most amenable to discontinuity detection.

If we proceed with the assumption that these experiments were p-hacked in some way, we would like to consider a way of handling these observations that would maximize our ability to detect this effect. Thinking through the dynamics of these experiments, it would seem that if p-hacking did play a role in determining their stopping times, the smallest p-value from each set of cotermi-

[21] We cannot say with certainty why firms are running experiments in this way. We have limited access to any detailed information about these experiments, other than a some metadata containing short titles and descriptions. In at least some of cases where experiments start/end at the same time, firms appear to be separating different device types into individual experiments; for example

nous experiments would play the largest role. Specifically, we imagine the most likely mechanism for p-hacking is as follows: a firm starts a group of experiments at the same time and continuously monitors the p-values from each test; upon observing one p-value dip below the critical 0.05 threshold, the firm then stops all experiments in the group. In this scenario, the p-values for the other experiments in the group will be above 0.05. We argue ignoring all but the smallest p-value in each set of coterminous experiments maximizes our ability to detect the effect of this type of group-level p-hacking. In the presence of p-hacking, dropping observations larger than the smallest p-value in a group should increase the mass of experiments to the left of the 0.05 threshold and decrease the mass to the right of this threshold. As such, even if our hypothesized mechanism is not precisely what determined the experiments' stopping times, this approach should be biased in favor of detecting a discontinuity.

When we apply the data filtering mechanism described above—i.e., identify groups of coterminous experiments and keep only the smallest p-value from each group—we are left with 2,049 observations.[22] Applying the same estimation strategy employed in Section 4 to this subsample of experiments, and then proceeding with the same discontinuity test developed earlier yields very similar results to our original analysis. For chosen bandwidth values of $h = 0.005, 0.01$ and 0.02, the p-values associated with the test for discontinuity at 0.05 are 0.95, 0.84, and 0.38, respectively. As with our prior specifications, we find no evidence for the hypothesized discontinuity which should be present if a portion of the experiments in our dataset were p-hacked.

7 Discussion & Conclusion

Accepting the apparent conclusion from our analysis—that firms in our sample engaged in little to no p-hacking behavior—prompts us to consider why a phenomenon that is widespread in other settings does not occur in ours. While we cannot claim to have proven this definitively, our findings are consistent with at least two possible explanations. One is that analysts at the firms in our sample are statistically sophisticated enough to know that continuously monitoring their p-values is poor research practice. Our results would also be consistent with the theory that private incentives of experimenters in economic settings are more aligned with the truth than those in other settings. Said differently, the consequences of p-hacking may be more salient or more severe for managerial decision makers than for academic researchers.

It is also useful to contextualize our findings in light of the previous work of Berman et al. (2018),

[22] We consider experiments by the same firm that both start and end within 5 minutes of each other to be effectively coterminous; varying this threshold has no material effect on subsequent conclusions.

who studied p-hacking in an ostensibly similar context. A careful comparison may provide some clues as to why our results differ and bring some clarity about the limits of our findings. To begin, we highlight several factors that appear to differ between our two studies. First, Berman et al. (2018) studies data from Optimizely which, at least during a part of their sample period, had a free tier for customers. The platform our data come from has never had a free tier, meaning that all firms in our sample paid a non-trivial monthly fee to use the testing platform's services. While it is not clear what fraction of experiments in their sample come from free-users, it appears possible that inherent differences in firm size, budget, and perhaps statistical sophistication are the cause of our discrepant findings. There may also be an economic explanation, whereby firms that have paid for a testing service are more incentivized to use it "properly". Further, the primary outcome metric analyzed by Berman et al. (2018) is "engagement"—which is defined merely as whether a user clicked anywhere on the page in their browsing session, and is the default reporting metric on Optimizely's platform. Recall that the default metric in our context is *conversion*, i.e., whether a user actually bought something during their session. Perhaps the simplest reason why this matters is that baseline conversion rates are much lower than baseline engagement rates. All else being equal, a proportional variable with a higher base rate will achieve statistical significance more often, meaning it would be easier to p-hack this variable. Another possibility is that because conversions are more closely translated into economic value, firms in our sample had stronger incentive to interpret the results of their experiments judiciously. A false positive for conversion rates would be much more costly than a false positive for engagement rates.

While a full accounting for the difference in our findings will certainly require more research, we can at least use this discussion to outline the conditions in which we expect our results to be more relevant than that of other work. Specifically, this would include A/B tests among firms that are mature enough to pay for experimentation software who are analyzing conversion rates and other similarly high-value, instrumental outcome metrics.

In conclusion, we highlight several managerial implications of our research project for both testing practitioners and experimentation platforms. First, our findings provide some evidence that p-hacking is not the default, inevitable outcome of providing experimenters with access to continuous data streams of experimental results—the current, standard industry practice in A/B testing. While many articles have been written about the perils of using A/B testing in managerial decision making, our results provide at least one counterpoint to this narrative. In the context of our data sample of medium to large e-commerce companies, the typical A/B test does not appear to

be the result of a contaminated data gathering process. Instead, we found that typical practices at the average firm are not obviously problematic, which may give managers a measure of confidence in using and interpreting A/B test results to inform consequential decisions. Similarly, these findings suggest that despite widespread concern about p-hacking, it may actually not be necessary to expend any resources attempting to fix a problem that may not exist. A reasonable conclusion from our research is that *before* making dramatic changes to experimentation practices—whether it be adopting a new statistical framework or retraining marketers on best practices—it is prudent for firms to determine whether p-hacking is indeed a problem for them in the first place. This is true for individual practitioners, experimentation teams at larger companies, and testing platforms themselves. Assuming firms have access to data on a reasonable number of prior experiments, the methods described in this paper can be used to investigate p-hacking behavior among proprietary datasets.

While more research on the phenomenon of p-hacking in A/B testing is clearly necessary, we believe this project—particularly since it appears to provide evidence against the prevailing narrative on the subject—serves as an important contribution to the academic literature on and industrial practice of digital experimentation.

References

Abadie, A. (2018). Statistical non-significance in empirical economics. Technical report, National Bureau of Economic Research.

Abhishek, V. and Mannor, S. (2017). A nonparametric sequential test for online randomized experiments. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 610–616. International World Wide Web Conferences Steering Committee.

Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C.-K., Prolla, T. A., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, 39(1):1–20.

Aquilonius, B. C. and Brenner, M. E. (2015). Students' resasoning about p-values. *Statistics Education Research Journal*, 14(2).

Aral, S., Brynjolfsson, E., and Wu, L. (2012). Three-way complementarities: Performance pay, human resource analytics, and information technology. *Management Science*, 58(5):913–931.

Azevedo, E. M., Alex, D., Montiel Olea, J., Rao, J. M., and Weyl, E. G. (2018). A/B testing.

Baker, G. P. (1992). Incentive contracts and performance measurement. *Journal of political Economy*, 100(3):598–614.

Bellemare, M. F. and Bloem, J. R. (2017). Experimental Conversations: Perspectives on Randomized Trials in Development Economics, by TN Ogden.

Berman, R., Pekelis, L., Scott, A., and Van den Bulte, C. (2018). p-Hacking and False Discovery in A/B Testing. *Available at SSRN: https://ssrn.com/abstract=3204791*.

Borden, P. (2014). How Optimizely (Almost) Got Me Fired. http://blog.sumall.com/journal/optimizely-got-me-fired.html. (Accessed on 08/17/2017).

Bothwell, L. E. and Podolsky, S. H. (2016). The emergence of the randomized, controlled trial. *New England Journal of Medicine*, 375(6):501–504.

Bourel, M. and Ghattas, B. (2012). Aggregating density estimators: an empirical study. *arXiv preprint arXiv:1207.4959*.

Brynjolfsson, E. and Hitt, L. (1996). Paradox lost? Firm-level evidence on the returns to information systems spending. *Management science*, 42(4):541–558.

Brynjolfsson, E., Hitt, L. M., and Kim, H. H. (2011). Strength in numbers: How does data-driven decisionmaking affect firm performance? *Available at SSRN 1819486*.

Brynjolfsson, E. and McElheran, K. (2016). The rapid adoption of data-driven decision-making. *American Economic Review*, 106(5):133–39.

BuiltWith (2019). A/B Testing Usage Distribution in the Top 1 Million Sites. https://web.archive.org/web/20190717062204/https://trends.builtwith.com/analytics/a-b-testing. (Accessed on 07/17/2019).

Cai, T. T. and Sun, W. (2017). Large-scale global and simultaneous inference: estimation and testing in very high dimensions. *Annual Review of Economics*, 9:411–439.

Carp, J. (2012). On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. *Frontiers in neuroscience*, 6:149.

Cattaneo, M. D., Jansson, M., and Ma, X. (2018a). Manipulation testing based on density discontinuity. *The Stata Journal*, 18(1):234–261.

Cattaneo, M. D., Jansson, M., and Ma, X. (2018b). Simple local polynomial density estimators. *arXiv preprint arXiv:1811.11512*.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, 83(403):596–610.

David, H. A. and Nagaraja, H. N. (2004). Order statistics. *Encyclopedia of Statistical Sciences*.

de Winter, J. C. and Dodou, D. (2015). A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ*, 3:e733.

Deng, A., Lu, J., and Chen, S. (2016). Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, pages 243–252. IEEE.

Dougherty, E. R. (2008). On the epistemological crisis in genomics. *Current Genomics*, 9(2):69–79.

Draper, P. (2016). The Fatal Flaw of A/B Tests: Peeking. https://www.lucidchart.com/blog/the-fatal-flaw-of-ab-tests-peeking. (Accessed on 06/11/2019).

Dreber, A. and Johannesson, M. (2019). Statistical Significance and the Replication Crisis in the Social Sciences. In *Oxford Research Encyclopedia of Economics and Finance*.

Earp, B. D. and Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in psychology*, 6:621.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Feng, E. (2017). Building an Intelligent Experimentation Platform with Uber Engineering. https://eng.uber.com/experimentation-platform/. (Accessed on 09/07/2018).

Franco, A., Malhotra, N., and Simonovits, G. (2016). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science*, 7(1):8–12.

Gartner (2019). A/B Testing Software. *G2 Track by Gartner*. https://www.g2.com/categories/a-b-testing.

Gelman, A. and Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, 37(3):447–456.

Gelman, A. and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*.

Gelman, A. and Loken, E. (2016). The statistical crisis in science. *The best writing on mathematics*, 2015:305.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, 31(4):337–350.

Gronau, Q. F., Duizer, M., Bakker, M., and Wagenmakers, E.-J. (2017). Bayesian mixture modeling of significant p values: A meta-analytic method to estimate the degree of contamination from H . *Journal of Experimental Psychology: General*, 146(9):1223.

Hamermesh, D. S. (2013). Six decades of top economics publishing: Who and how? *Journal of Economic Literature*, 51(1):162–172.

Hartgerink, C. H. (2017). Reanalyzing Head et al.(2015): investigating the robustness of widespread p-hacking. *PeerJ*, 5:e3068.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS biology*, 13(3):e1002106.

Heck, P. R., Chabris, C., Watts, D. J., and Meyer, M. N. (2019). Sometimes People Dislike Experiments More than They Dislike Their Worst Conditions: Within-Subjects Evidence for "Experiment Aversion" and the A/B Effect. *Available at PsyArxiv https://doi.org/10.31234/osf.io/jmxgc*.

Hern, A. (2014). Why Google has 200m reasons to put engineers over designers. *The Guardian*. https://www.theguardian.com/technology/2014/feb/05/why-google-engineers-designers.

Holmstrom, B. and Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *JL Econ. & Org.*, 7:24.

Hubbard, R. (2011). The widespread misinterpretation of p-values as error probabilities. *Journal of Applied Statistics*, 38(11):2617–2626.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *The Journal of Experimental Education*, 61(4):317–333.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8):e124.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5):524–532. PMID: 22508865.

Kleven, H. J. (2018). Language Trends in Public Economics.

Kohavi, R. (2019). History of Controlled Experimentation. *Available at experimentationguide.com/history/*.

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., and Pohlmann, N. (2013). Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1168–1176. ACM.

Kohavi, R. and Longbotham, R. (2017). Online controlled experiments and a/b testing. In *Encyclopedia of machine learning and data mining*, pages 922–929. Springer.

Koning, R., Hasan, S., and Chatterji, A. (2019). Experimentation and Startup Performance: Evidence from A/B testing. Working Paper 26278, National Bureau of Economic Research.

Krawczyk, M. (2015). The search for significance: a few peculiarities in the distribution of P values in experimental psychology literature. *PloS one*, 10(6):e0127872.

Leggett, N. C., Thomas, N. A., Loetscher, T., and Nicholls, M. E. (2013). The life of p:"Just significant" results are on the rise. *The Quarterly Journal of Experimental Psychology*, 66(12):2303–2309.

Liu, C. and Chamberlain, B. P. (2018). Online Controlled Experiments for Personalised e-Commerce Strategies: Design, Challenges, and Pitfalls. *arXiv preprint arXiv:1803.06258*.

Lu, L. (2016). Power, minimal detectable effect, and bucket size estimation in A/B tests.

https://blog.twitter.com/engineering/en_us/a/2016/power-minimal-detectable-effect-and-bucket-size-estimation-in-ab-tests.html. (Accessed on 09/07/2018).

Masicampo, E. and Lalande, D. R. (2012). A peculiar prevalence of p values just below. 05. *The Quarterly Journal of Experimental Psychology*, 65(11):2271–2279.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., and Barton, D. (2012). Big data: the management revolution. *Harvard business review*, 90(10):60–68.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, 142(2):698–714.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman and Hall/CRC.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, 6:355–378.

Miller, A. P. and Hosanagar, K. (2018). An Exploratory Meta-analysis of Empirical E-commerce A/B Testing Strategies. *Working Paper*.

Miller, E. (2010). How Not To Run an A/B Test. http://www.evanmiller.org/how-not-to-run-an-ab-test.html. (Accessed on 08/17/2017).

Mislavsky, R., Dietvorst, B., and Simonsohn, U. (2019). Critical Condition: People Don't Dislike A Corporate Experiment More than They Dislike Its Worst Condition. *Available at SSRN 3288076*.

Mitchell, W. J., Inouye, A. S., and Blumenthal, M. S. (2003). *Beyond productivity: Information technology, innovation, and creativity*. National Academies Press.

Nettleton, D., Hwang, J. G., Caldo, R. A., and Wise, R. P. (2006). Estimating the number of true null hypotheses from a histogram of p values. *Journal of agricultural, biological, and environmental statistics*, 11(3):337.

Otsu, T., Xu, K.-L., and Matsushita, Y. (2013). Estimation and inference of discontinuity in density. *Journal of Business & Economic Statistics*, 31(4):507–524.

Overgoor, J. (2014). Experiments at Airbnb. https://medium.com/airbnb-engineering/experiments-at-airbnb-e2db3abf39e7. (Accessed on 09/07/2018).

Parker, R. and Rothenberg, R. (1988). Identifying important results from multiple statistical tests. *Statistics in medicine*, 7(10):1031–1043.

Pekelis, L., Walsh, D., and Johari, R. (2015). Optimizely's New Stats Engine (White Paper).

Perneger, T. V. and Combescure, C. (2017). The distribution of P-values in medical research articles suggested selective reporting associated with statistical significance. *Journal of clinical epidemiology*, 87:70–77.

Peysakhovich, A. and Eckles, D. (2018). Learning causal effects from many randomized experiments using regularized instrumental variables. In *Proceedings of the 2018 World Wide Web Conference*, pages 699–707. International World Wide Web Conferences Steering Committee.

Pounds, S. and Morris, S. W. (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(10):1236–1242.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3):638.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, pages 65–78.

Sanders, M. and Halpern, D. (2014). Nudge unit: our quiet revolution is putting evidence at heart of government. *The Guardian*.

Schneider, J. W. (2015). Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics*, 102(1):411–432.

Sijtsma, K. (2016). Playing with data—or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika*, 81(1):1–15.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366.

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2):534.

Smith, K. N., Makel, M. C., and Plucker, J. (2019). Improving Research in Educational Psychology, Psychology, and the Social Sciences. *PsyArXiv*. doi:10.31234/osf.io/ams9e.

Spiess, J. (2018). Optimal estimation when researcher and social preferences are misaligned.

Szucs, D. (2016). A tutorial on hunting statistical significance by chasing N. *Frontiers in psychology*, 7:1444.

Tang, Y., Ghosal, S., and Roy, A. (2007). Nonparametric Bayesian estimation of positive false discovery rates. *Biometrics*, 63(4):1126–1134.

Tauber, S. (1963). On multinomial coefficients. *The American Mathematical Monthly*, 70(10):1058–1063.

Tsybakov, A. B. (2009). Springer Series in Statistics.

Virzi, A. M. (2018). A/B Testing in Marketing: The Customer's Always Right By Anna Maria Virzi. *Gartner for Marketers*. https://blogs.gartner.com/anna-maria-virzi/2018/02/08/ab-testing-in-marketing-the-customers-always-right/.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The annals of mathematical statistics*, 16(2):117–186.

Ware, J. J. and Munafò, M. R. (2015). Significance chasing in research practice: causes, consequences and possible solutions. *Addiction*, 110(1):4–8.

Woolf, P. K. (1986). Pressure to publish and fraud in science. *Annals of Internal Medicine*, 104(2):254–256.

APPENDICES

A The minimum order statistic for the beta-uniform mixture model

For the sake of generality we derive our proofs below with an arbitrary number of mixture components K and sample sizes M. In the empirical model described in the main body of the text, the number of components $K = 3$ and the number of draws (representing the number of different variables a firm may be simultaneously monitoring) $M = 8$.

PROPOSITION 1. *Assume X is random variable with density function defined by a k-component beta-uniform mixture model. Formally, $X \sim F$, with density $DF = f$ given by*

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x) \tag{9}$$

where f_1 is the uniform density and the density for each $k > 1$ is defined as that of a Beta-distributed random variable:

$$f_k(x) = \begin{cases} 1, & \text{for } k = 1 \\ \frac{1}{\mathrm{B}(\alpha_k, \beta_k)} x^{\alpha_k - 1}(1-x)^{\beta_k - 1}, & \text{for } k \geq 2 \end{cases}$$

Denote each component's marginal distribution function F_k so that $F(x) = \sum \pi_k F_k(x)$. Then the density of the first order statistic for M i.i.d. draws of X, $\{X_m \mid X_m \sim X\}$, denoted

$$X_{(1)}^{M} = \min\{X_1, X_2, \dots, X_M\} ,$$

is given by

$$f_{(1)}^{M}(x) = \sum_{j_0 + j_1 + \cdots + j_K = M} (-1)^{1+M-j_0} \binom{M}{j_0, j_1, \dots, j_K} \left(\prod_{k=1}^{K} (\pi_k F_k(x))^{j_k} \right) \left(\sum_{k=1}^{K} \frac{j_k f_k(x)}{F_k(x)} \right) \tag{10}$$

Proof. Using the formula for the CDF of the mininum order statistic (David and Nagaraja, 2004), we can derive:

$$\begin{aligned} F_{(1)}(x) &= 1 - [1 - F(x)]^M \\ &= 1 - \left[1 - \sum_{k=1}^{K} \pi_k F_k(x) \right]^M \\ &= 1 - \left[1 + \sum_{k=1}^{K} -\pi_k F_k(x) \right]^M \\ &= 1 - \left[\sum_{k=0}^{K} g_k(x) \right]^M \end{aligned} \tag{11}$$

where we have defined

$$g_k(x) = \begin{cases} 1, & \text{for } k = 0 \\ -\pi_k F_k(x), & \text{for } k \geq 1 \end{cases} \tag{12}$$

We can now use the multinomial expansion formula (Tauber, 1963) to rewrite Eq. (11):

$$= 1 - \left[\sum_{j_0+j_1+\cdots+j_K=M} \binom{M}{j_0, j_1, \ldots, j_K} \prod_{k=0}^{K} g_k(x)^{j_k} \right] \tag{13}$$

The expression in parentheses used above is the multinomial coefficient given by:

$$\binom{M}{j_0, j_1, \ldots, j_K} = \frac{M!}{j_0! j_1! \cdots j_K!}$$

Because $g_0(x) = 1$, we can drop the first $k = 0$ term in the product operator, so the index runs from $k = 1, \ldots, K$.

To obtain an expression for the density of $X_{(1)}^M$, we merely need to differentiate this expression for $F_{(1)}$:

$$\begin{aligned} f_{(1)}(x) &= \frac{d}{dx}\left\{F_{(1)}(x)\right\} \\ &= \frac{d}{dx}\left\{1 - \left[\sum_{j_0+j_1+\cdots+j_K=M} \binom{M}{j_0, j_1, \ldots, j_K} \prod_{k=1}^{K} g_k(x)^{j_k} \right]\right\} \\ &= -\sum_{j_0+j_1+\cdots+j_K=M} \binom{M}{j_0, j_1, \ldots, j_K} \frac{d}{dx}\left\{\prod_{k=1}^{K} g_k(x)^{j_k}\right\} \\ &= -\sum_{j_0+j_1+\cdots+j_K=M} \binom{M}{j_0, j_1, \ldots, j_K} \frac{d}{dx}\left\{\prod_{k=1}^{K} u_k(x)\right\} \end{aligned} \tag{14}$$

where in the last equality we have substituted

$$u_k(x) = g_k(x)^{j_k}$$

By applying a generalized product rule (differentiation of products involving more than two terms), we can rewrite the differentiated term as:

$$\begin{aligned} \frac{d}{dx}\left\{\prod_{k=1}^{K} u_k\right\} &= \sum_{k=0}^{K} \left(u_k' \prod_{\substack{0 \leq t \leq K \\ k \neq t}} u_t \right) \\ &= \left(\prod_{k=1}^{K} u_k\right)\left(\sum_{k=1}^{K} \frac{u_k'}{u_k}\right) \end{aligned} \tag{15}$$

Substituting back in the definition of u_k in the expression in the last summation over $\frac{u_k'}{u_k}$, we

obtain

$$\begin{aligned}\frac{u_k(x)'}{u_k(x)} &= \frac{(g_k(x)^{j_k})'}{g_k(x)^{j_k}} \\ &= \frac{j_k g_k(x)^{j_k-1} g_k'(x)}{g_k(x)^{j_k}} \\ &= \frac{j_k g_k'(x)}{g_k(x)}\end{aligned} \tag{16}$$

Recalling the definition of g_k, we can rewrite the minimum order statistic in Eq. (14) as

$$\begin{aligned}f_{(1)}^M(x) &= -\sum_{j_0+j_1+\cdots+j_K=M} \binom{M}{j_0,j_1,\ldots,j_K} \left(\prod_{k=1}^K g_k(x)^{j_k}\right) \left(\frac{j_k g_k'(x)}{g_k(x)}\right) \\ &= -\sum_{j_0+j_1+\cdots+j_K=M} \binom{M}{j_0,j_1,\ldots,j_K} \left(\prod_{k=1}^K (-\pi_k F_k(x))^{j_k}\right) \left(\sum_{k=1}^K \frac{j_k(-\pi_k f_k(x))}{-\pi_k F_k(x)}\right) \\ &= -\sum_{j_0+j_1+\cdots+j_K=M} \binom{M}{j_0,j_1,\ldots,j_K} \left(\prod_{k=1}^K (-\pi_k F_k(x))^{j_k}\right) \left(\sum_{k=1}^K \frac{j_k f_k(x)}{F_k(x)}\right) \\ &= -\sum_{j_0+j_1+\cdots+j_K=M} \binom{M}{j_0,j_1,\ldots,j_K} \left(\prod_{k=1}^K (-1)^{j_k}\right) \left(\prod_{k=1}^K (\pi_k F_k(x))^{j_k}\right) \left(\sum_{k=1}^K \frac{j_k f_k(x)}{F_k(x)}\right) \\ &= -\sum_{j_0+j_1+\cdots+j_K=M} \binom{M}{j_0,j_1,\ldots,j_K} \left((-1)^{\sum_{k=1}^K j_k}\right) \left(\prod_{k=1}^K (\pi_k F_k(x))^{j_k}\right) \left(\sum_{k=1}^K \frac{j_k f_k(x)}{F_k(x)}\right) \\ &= \sum_{j_0+j_1+\cdots+j_K=M} (-1)^{1+M-j_0} \binom{M}{j_0,j_1,\ldots,j_K} \left(\prod_{k=1}^K (\pi_k F_k(x))^{j_k}\right) \left(\sum_{k=1}^K \frac{j_k f_k(x)}{F_k(x)}\right)\end{aligned} \tag{17}$$

This demonstrates the desired result.

■

The density derived above can be used in likelihood based estimation techniques for beta-uniform mixture models (as is done in the text).

B Power calculation with counterfactual simulation

To calculate the power of our test for a given sample size, N, and effect size q, we must integrate out the sampling variation associated with our simulation procedure. Sampling variation is introduced in two phases of our simulation: first, when selecting which experiments will be p-hacked and, in the case that the sample size N differs from that of our dataset, which experiments are selected into the simulated sample. In this section, we provide pseudo-code for our technique (Algorithm 1) which uses a bootstrap aggregation procedure for effectively integrating out this variation, leaving us with the average power of our procedure across independent trials.

Algorithm 1: Power calcuation pseudocode

Input: $N \in \mathbb{N}_+$, sample size (number of experiments);
$q \in [0, 1]$, effect size (proportion of experiments p-hacked);
Output: Rejection rate (power, $1 - \beta$) of discontinuity test at chosen α level, averaged by bootstrap
Fixed variables:
E, dataset of experiments, each experiment consists of a daily time-series of empirically observed p-values;
$\alpha = 0.05$, false positive rate;
$B >> 0$, bootstrap sample size ;
Notation:
T_i, total number of days in experiment i;
p_{it}, empirically observed p-value of experiment i calculated on day $t \leq T_i$;
p_i^*, terminal p-value observed for experiment i during simulation;
$[Z]$, the set of integers $\{1, \ldots, Z\}$ for $Z \in \mathbb{N}_+$;
$\mathscr{S}(\cdot)$, test statistic for discontinuity, as described in Section 4 (returns p-value);
$\mathscr{P}$, set of p-values for statistical test, observed across bootstrap samples

```
1  function PowerCalculation(q, N)
2      P ← {}
3      for bootstrap b ∈ [B] do
4          E_N ← random sample of size N from E                    with replacement
5          P_b ← {}
6          H_b ← random sample of proportion q from E_N
7          for experiment i ∈ E_N do
8              if i ∈ H_b then
9                  for day t ∈ [T_i] do
10                     if p_it < 0.05 then
11                         p_i* ← p_it                             set to p-hacked value
12                         break loop; go to 15
13             else
14                 p_i* ← p_iT_i                                   use empirically observed terminal p-value
15             P_b ← {p_i*} ∪ P_b                                  add simulated p-value to observation set
16         π_b ← S(P_b)                                            apply statistical test for discontinuity to p-hacked p-values
17         P ← {π_b} ∪ P                                           collect p-values of discontinuity test
18     R ← {1[p < α] for p ∈ P}                                    indicator set of rejected tests
19     power ← ∑R/B                                                average rejection proportion across bootstraps
```

C Selecting bandwidth parameter h

The primary bandwidth used for our testing procedure in the body of the text is fixed at $h = 0.01$. While this may appear arbitrary, we have attempted to use more data-adaptive methods for selecting the bandwidth value. While these methods may be appropriate to use with large N, our findings suggest that—at our current sample size of less than 2,500 observations—data-driven bandwidth selection exhibits small-sample effects that make such an approach less than ideal for the problem of discontinuity detection. To see why this is the case, we describe a straightforward comparison of data-adaptive and fixed bandwidth methods below.

We must first select a method for determining a data-adaptive optimal bandwidth. We argue the goals of our test align well with the goals of selecting an optimal bandwidth of a histogram density estimator. Such procedures are designed to minimize the L^2 risk of an estimator, i.e., its mean integrated square error:

$$MISE(\hat{f}) = \mathbf{E} \int \left(\hat{f}(x) - f(x) \right)^2 dx \, .$$

The bandwidth selection problem can then be formally expressed as:

$$h^* = \arg\max_h MISE(\hat{f}_h)$$

For the family of *histogram estimators*, one approach for solving for this optimization problem empirically is leave-one-out cross-validation (LOOCV) Rudemo (1982). This method has a convenient analytical formula for the histogram estimator's mean integrated squared error, a standard loss function in the density estimation literature:

$$MISE(h) = \frac{2}{(n-1)h} - \frac{n+1}{n^2(n-1)h} \sum_k N_k^2$$

where n is the total number of observations and N_k is the number of observations in the k-th histogram bin (Tsybakov, 2009).

We evaluated the performance of our test procedure using both the LOOCV bandwidth method and the fixed bandwidth method (as done in the main text). By calculating the same power contour plot shown in Figure 8a, but substituting these different methods of choosing h. We are able to estimate the power of each approach empirically. Since we, as the researchers designing the counterfactual power simulations known the effect size q is always positive (i.e., we p-hacked some portion of our experiments in the simulations), calculating the power of our test amounts to simply estimating the fraction of times that our test rejects the null hypothesis at the chosen $\alpha = 0.05$ level. Figure 11 shows the results of these calculations for scenarios in which we used either the

fixed or adaptive values of selecting the optimal h. The first row contains the power contours for our testing procedure with h taking on the fixed values of $0.005, 0.01$ and 0.02, respectively. In the second row, we use the data-adaptive LOOCV method described above to estimate the optimal $\hat{h}$ separately for each run of the simulation; we show the results of our power analysis when the bandwidth of the testing procedure is set to $\hat{h}/2, \hat{h}$, and $2\hat{h}$. The third row of Figure 11 shows the average value of the calculated bandwidth used at each point in the grid.

In contrast to the top row in which h is a fixed value, the power contours for testing procedures with adaptive h values are highly irregular. Note that a good statistical test will consistently increase in power as either the sample size or effect size increases (as is the case for the tests with fixed h). However, for the tests with adaptive h values, statistical power exhibits non-monotonic behavior with respect to sample size, sometimes increasing and sometimes decreasing as sample size grows. While these irregularties appear to mellow out for large samples sizes (near $N = 10,000$), the size of our empirical sample is only around 2,500. As such, it appears imprudent to ignore the small-sample effects of the data-adaptive estimators. For this reason, we choose to use fixed values of h, with $h = 0.01$ as our primary bandwidth parameter, for analysis of our dataset. However, because we have reported statistics for all values $h = 0.005, 0.01$, and 0.02 throughout this paper, one can see that our results to not appear particularly sensitive to this choice.

Figure 11: Effects of fixed vs. adaptive bandwidth values on statistical power