

An investigation of p -hacking in e-commerce A/B testing

Alex P. Miller
University of Southern California
Marshall School of Business
alex.miller@marshall.usc.edu

Kartik Hosanagar
The Wharton School
University of Pennsylvania
kartikh@wharton.upenn.edu

Forthcoming in *Information Systems Research*

Abstract

In recent years, randomized experiments (or “A/B tests”) have become commonplace in many industrial settings as managers increasingly seek the aid of scientific rigor in their decision-making. However, just as this practice has proliferated among firms, the problem of p -hacking—whereby experimenters adjust their sample size or try several statistical analyses until they find one that produces a statistically significant p -value—has emerged as a prevalent concern in the scientific community. Notably, many commentators have highlighted how A/B testing software enables and may even encourage p -hacking behavior. To investigate this phenomenon, we analyze the prevalence of p -hacking in a primary sample of 2,270 experiments conducted by 242 firms on a large U.S.-based e-commerce A/B testing platform. Using multiple statistical techniques—including a novel approach we call the *asymmetric caliper test*—we analyze the p -values corresponding to each experiment’s designated target metric across multiple significance thresholds. Our findings reveal essentially no evidence for p -hacking in our data. In an extended sample that examines p -hacking across all outcome metrics (encompassing more than 16,000 p -values in total), we similarly observe no evidence of p -hacking behavior. We use simulations to determine that if a modest effect of p -hacking were present in our dataset, our methods would have the power to detect it at our current sample size. We contrast our results with the prevalence of p -hacking in academic contexts and discuss a number of possible factors explaining the divergent results, highlighting the potential roles of organizational learning and economic incentives.

Keywords: p -hacking, A/B testing, Data-driven decision-making, Electronic commerce, Mixture models

Notes: Thank you to the editors and anonymous review team. Thank you to participants and discussants at the Conference on Digital Experimentation (CODE), the Workshop on Information Systems & Economics (WISE), and the Conference on Information Systems & Technologies (CIST). The authors are grateful to the Baker Retailing Center and the Mack Institute for Innovation Management at the University of Pennsylvania for helping fund this work.

1 Introduction.

In the first half of the 20th century, a British industrial researcher named Ronald Fisher published several seminal works that would go on to profoundly shape the field of statistical science. Fisher’s contributions, particularly his development and advocacy of “ p -values”, significance testing, and experimental design, formalized the use of statistical methods for empirical research. His emphasis on randomization and his pioneering work on the analysis of variance revolutionized experimental methodology across various disciplines in both practical industry and academic science. Over time, Fisher’s approaches have become fundamental to decision-making processes in fields ranging from agriculture and economics to psychology and medicine (Hamermesh, 2013, Kleven, 2018, Kohavi, 2019).

As statistical methods became more established throughout the 20th century, the practice of data-driven decision-making gained prominence (Brynjolfsson and McElheran, 2016, Schneider, 2015). This is particularly true in *digital* business, where the marginal cost of both service innovation and delivery is relatively low. Dozens of software-as-a-service (SaaS) platforms have launched in recent years that enable firms to run experiments on their websites and apps, sometimes at no cost. A growing body of research on digital experimentation—now widely known as “A/B testing” among software, marketing, and e-commerce companies—has emerged, including work from academics in many fields such as economics, computer science, statistics, information systems, and marketing (Azevedo et al., 2018, Kohavi et al., 2013, Liu and Chamberlain, 2018). In industrial contexts, A/B testing has emerged as a vital 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). While the largest technology companies have used online experiments for decades, use of A/B testing is now growing among firms of all types. By some estimates, more than 40% of the top 10,000 websites by traffic are using third-party technologies with A/B testing capabilities (BuiltWith, 2019).¹ Further, a 2018 survey of more than 200 companies with at least \$500 million in annual revenue indicated that 74% of responding firms indicated they either already use or plan to use A/B testing in the near future (Virzi, 2018).

Just as randomized experiments have proliferated in the digital economy, there has been increasing interest in the downsides of the conventional statistical methods commonly used to analyze these experiments. Much of this literature focuses on the shortcomings of null hypothesis signifi-

¹This figure does not account for internally developed experimentation tools, which are known to be widely used at large technology companies.

cance testing (NHST), the prevailing scientific approach for assessing the long-term (frequentist) error rates in decisions based on p -values. In common statistical practice, NHST categorizes results as “insignificant” when p -values exceed 0.05 and “significant” when p -values fall below 0.05. This rule-of-thumb can be traced back to Ronald Fisher’s 1925 work, *Statistical Methods for Research Workers*, wherein Fisher suggested it would be “convenient to take” the $p=0.05$ threshold “as a limit in judging whether a deviation is to be considered significant or not” (Fisher, 1925).

As empirical science and statistics evolved and became institutionalized throughout the 20th century, this simple rule-of-thumb gradually solidified into a widespread convention for evaluating statistical evidence, with the $p=0.05$ threshold often being treated as a definitive cutoff.² In recent decades, however, the scientific community has increasingly scrutinized the utility, applicability, and consequences of widespread reliance on significance testing that dichotomizes results based on p -values in this way. For example, it has been argued that published scientific 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 for “significant” results induced by the peer-review and publication process (Ioannidis, 2005, Rosenthal, 1979). In addition to the 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, Simmons et al., 2011). This phenomenon by which analysts change their data sampling procedures or statistical techniques to obtain “significant” results has come to be known as “ p -hacking” (Simmons et al., 2013). 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” (Dreber and Johannesson, 2019, Earp and Trafimow, 2015).

Despite this crisis, NHST and p -values have come to predominate much of the statistical software used throughout the A/B testing industry. However, given that A/B testing results are reported in near real-time, this enables a potentially pernicious form of p -hacking referred to as continuous monitoring (or “optional stopping”), whereby experimenters regularly check a test’s p -value and only end an experiment when a sufficiently small 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. While analysts at larger and more technical organizations may be aware of the pitfalls of continuous monitoring,

²See 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. Also see Leahey (2005), who documents how the adoption of statistical reporting practices among prestigious editors and institutions led to the proliferation of significance testing in social science.

the rise of low-cost A/B testing platforms has significantly increased the ability of firms of all sizes to run experiments.³ Indeed, A/B testing software is specifically marketed to facilitate to use of randomized experiments among firms without the know-how to develop their own technology or the statistical expertise to analyze experimental results. Given that academic researchers, often with doctoral degrees and graduate training in statistics, have been known to engage in p -hacking behavior (Brodeur et al., 2020, 2023b, Szucs, 2016), it is reasonable to ask whether analysts in corporate environments using similar statistical techniques make similar methodological errors.

Answering this question can have far-reaching implications for how researchers and managers understand the value of A/B testing and statistical methodologies in both academic and industrial contexts. If p -hacking behavior is a widespread phenomenon on A/B testing platforms, many firms would be justified in reevaluating how they use experiments for business decisions. On the other hand, if p -hacking is less common in industrial settings than academic settings, this would be an interesting result in multiple regards. First, such a result would suggest that digital experiments in real-world settings may not be as fraught with error as the existing literature on the subject might suggest. This may increase the confidence that managers and executives have about the use of statistics and A/B testing for making business decisions more generally. Further, if p -hacking is absent in the industrial context, this would put in stark contrast the ubiquity of p -hacking in academic science, potentially highlighting the need for more research studying precisely if and how contextual factors drive p -hacking behavior. In either case, we argue there is scientific and managerial value in a credible empirical analysis on the incidence of p -hacking on digital testing platforms.

This discussion motivates the current project, in which we set out to investigate the prevalence of p -hacking behavior on a large A/B testing platform. We proceed by analyzing the distribution of p -values from a sample of A/B tests conducted by e-commerce merchants. Our analysis will exploit the fact that, if experimenters do consistently stop their tests right when their p -values reach 0.05, or gather extra data for experiments that have yet to reach this value, we would expect to see a jump in the number of p -values observed right below this threshold. To detect this effect, we use both existing statistical techniques for detecting such discontinuities and also develop a new method that is more robust to nuisance parameters than existing approaches found in related literature. In our primary analysis, we apply these techniques to the distribution of p -values from 2,270 experiments conducted by 242 firms and find little to no evidence for the incidence of p -hacking in our dataset. In secondary analyses, this null result remains robust to several different assumptions

³At larger organizations, a significant amount of resources has been allocated to studying and ameliorating the problems created by continuous monitoring; research labs at Microsoft, Walmart, 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).

about firm behavior. 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. In sum, this research makes valuable contributions by presenting both a robust approach to estimating discontinuities in p -value distributions and credible empirical evidence about how real-world firms use and deploy statistical tools for managerial decision-making.

2 Background & Motivation.

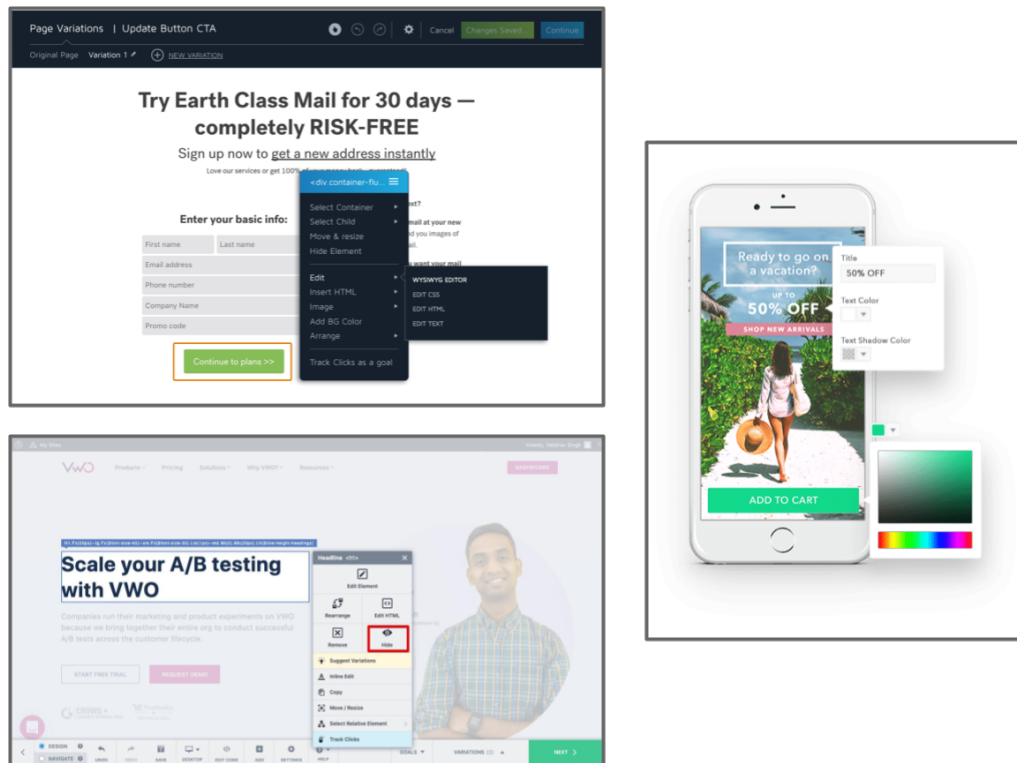
2.1 Primer on A/B testing.

2.1.1 Evolution of A/B testing and the role of experimentation platforms. Prior to the modern IT revolution, A/B testing was an *ad hoc* procedure largely practiced within isolated teams for specialized purposes. Academic and practitioner-oriented work documents the use of random experimentation in advertising, sales, and direct mail campaigns at large enterprises (Chickering and Heckerman, 2000, Tang et al., 2010, Warwick, 2003). However, as much of the economy became digitized throughout the early 2000s, the costs associated with implementing, measuring, and analyzing experiments dropped significantly. It was around this time that large organizations such as Google and Microsoft started developing standardized software tools and organizational practices around A/B testing, which ultimately ended up being core to how many decisions at these companies got made (Kohavi and Thomke, 2017). As legacy enterprises started adapting to the information economy and new digital-native companies emerged, the capital developed by these successful IT companies—both human and technological—began to influence organizational practices in the economy at large (Tambe and Hitt, 2014).

Within these broader shifts in the economy, the diffusion of A/B testing specifically was enabled by new platforms that began to offer “experimentation-as-a-service”. Around the turn of the 2010s, a number of these platforms emerged as both additional products offered by existing software companies (e.g., Google Optimize, Microsoft Azure, Adobe Test & Target, Mixpanel, HubSpot) and independent start-ups focused on A/B testing as a key value proposition (e.g., Optimizely, Visual Website Optimizer, AB Tasty, Experiment.ly, Split Optimizer). In addition to handling technical challenges with digital experimentation—such as user allocation, behavior tracking, and statistical reporting—a pivotal component to the commercial success of these platforms was their interactive “WYSIWYG” (what you see is what you get) website editors. This dramatically lowered the cost of developing new website variations for non-engineering teams, which was a crucial step in facilitating the spread of A/B testing across and within organizations of all sizes (Christian, 2012a). With the combined influence of successful tech companies that were known for relying on digital experiments and the marketing efforts of these new testing-focused platforms, A/B testing has since

developed over the past decade into a well-established practice in the management of many types of organizations (Christian, 2012b). Many executives, managers, consultants, and academics now recognize A/B testing as a critical tool in modern managerial decision-making (Thomke, 2020).

Figure 1: Examples of WYSIWYG editors used to create variations in A/B testing software



2.1.2 A/B testing in e-commerce. These developments have been particularly relevant to managers of retail e-commerce companies. Such managers face many decision problems for which A/B testing can be beneficial: interface design, marketing messaging, promotional strategies, product placements, and algorithmic recommendations can all dramatically affect consumer behavior. And, in contrast to other sectors that have to contend with complex statistical estimation issues (such as long sales cycles in enterprise software or network interference between users on social media platforms), classical conditions required for causal inference in experimental settings are typically satisfied in e-commerce settings by simply randomizing individual users in an experiment and measuring their short-term response. It further helps that, in e-commerce, the key indicators of consumer behavior that managers are primarily interested in include easily quantifiable metrics that can be measured within a single website session, such as purchase incidence and revenue. These conditions, along with advent of user-friendly experimentation tools described above, have led to the rapid adoption of A/B testing among e-commerce companies.

Notably, managers of online retail stores using third-party A/B testing software are likely to have backgrounds quite different from the earliest adopters of A/B testing practices. Those who

contributed to the development of internal A/B testing software at large technology companies are likely to have had specific expertise in engineering or statistics and analytics. However, managers of online retail companies may have earned their positions due to skills in other domains, such as brick-and-mortar retail, trend forecasting, fashion, and marketing. Relative to their tech-industry counterparts, these managers are more likely to learn about website development, statistics, and A/B testing in an *ad hoc* way throughout their job. While this is precisely the audience that A/B testing platforms seek to serve with their user-friendly tools, we believe the extent to which this lack of technical expertise affects the value of A/B testing and analytics software overall is an under-explored question in existing academic literature and a vital component of the motivation for the current research project.

2.2 A/B testing interfaces. An important challenge for developers of A/B testing tools, particularly given their goal to reach non-technical users, is deciding how to communicate results and concepts around statistical uncertainty. As the industry matured, there has been a notable degree of technological convergence, with a number of implicit standards in testing interfaces that have come to be expected by users. Essentially every testing tool allows users to view basic analytics associated with each arm in a given experiment; this typically includes the number of participants in each arm, along with key outcome metrics relevant to the particular experimental context. In many scenarios, this will include some form of “conversion rate”—the percentage of users that took a desired action (e.g., product purchase, account creation). When they are relevant, other metrics can also be included in a results dashboard, such as session length, pageviews, bounce rate, revenue, and load time. It is also common, though not universal, for platforms to show time-series charts of how various metrics have evolved over time. When it comes to directly comparing the performance of each arm in an A/B test, standard practice is to quantify how each experimental treatment differs from a selected “baseline” or “control” treatment along the metrics being measured. For each metric, it is common to report the measured effect size between each non-baseline treatment and the baseline treatment (sometimes described as the “lift”). Depending on the platform, this effect size will be reported as either a level (additive) or proportional (multiplicative) change. We have documented several A/B testing interfaces in Appendix A, where we provide various real-world examples of how different testing platforms operationalize these features (see Figure A.1).

Though it is by no means universal, the most common statistical paradigm seen in A/B testing software is based on classical null hypothesis significance testing. This is perhaps unsurprising given the predominance of NHST throughout much of statistical science. On the whole, there is no standard around exactly which test is used to calculate p -values, as there are many reasonable

choices for testing the difference between two proportions in the statistical literature (e.g., Z -test, binomial proportion test, χ^2 -test). There are also differing practices around whether one-sided or two-sided hypothesis tests are used, whether finite sample corrections are accounted for, or if variances between the two treatment populations are assumed to be equal or unequal.

A simple method used by many platforms (including the one studied in this project) is a two-sided Z -test for testing the difference in sample means, where it is assumed that sample sizes are large enough to invoke the Central Limit Theorem (CLT) so that the test statistic is normally distributed under the null hypothesis (Kohavi et al., 2009). In mathematical terms, the p -value for this test can be calculated by assuming one observes both the number of conversions (represented by c_t) and the total number of users (n_t) in each of the treatment conditions of a two-armed A/B test, where treatment is indicated by subscript $t \in \{a, b\}$. The observed effect size of an experiment (or lift) is then defined as the difference in mean conversion rates between the two treatment conditions: $\hat{\mu} = c_b/n_b - c_a/n_a$. The empirically observed Z -score can then be computed by calculating the standard error of the mean difference: $\hat{\sigma} = \sqrt{\hat{p}_a(1-\hat{p}_a)/n_a + \hat{p}_b(1-\hat{p}_b)/n_b}$ where $\hat{p}_t = c_t/n_t$. This standard error is then divided into the estimated mean, $\hat{\mu}$, to obtain the primary test statistic of interest, $Z := \hat{\mu}/\hat{\sigma}$.

To derive the p -value associated with this test, one must then compare this empirically observed statistic, Z , to the theoretical distribution for this statistic under the null hypothesis—i.e., the assumption that there is no difference in conversion rates between treatment arms. We denote this theoretical test statistic (the null test statistic) as z , whose distribution can be derived in closed form as the standard Gaussian (normal) distribution using the CLT. The p -value of an experiment is then defined as the theoretical probability that the null test statistic, z , would be as large or larger than the empirically observed test statistic, Z : $p\text{-value} := P[z \geq |Z| \mid z \sim \Phi] = 2(1 - \Phi(|Z|))$

Interestingly, almost no testing platform that relies on p -values of any kind describes them as such. Instead, a very common practice in the industry is to show analysts a value equal to $(1 - p) \times 100$, and to use a term of art, such as “confidence”, to signal how this number is to be interpreted. For example, a test with a p -value of 0.13 will be shown to analysts as “87% confidence”. Other phrases in the industry used for this quantity include “statistical significance” level or even the explicitly misleading phrase “chance to beat baseline”. Independent of how testing platforms compute p -values or describe them to end-users, another characteristic common to nearly all platforms is how they discretize statistical uncertainty. This practice likely originated as a consequence of both the prominent use of “significance” thresholds throughout the history of academic science, as well as managers’ inherent need to make binary “ship” or “don’t ship”

decisions about the interventions being tested on these platforms.

2.3 *p*-hacking in A/B testing.

2.3.1 Potential for p-hacking and existing evidence. To motivate our research questions, we highlight a few key factors from the preceding discussion on the A/B testing industry. First, we emphasize that modern A/B testing tools are explicitly designed to allow non-experts to run randomized experiments for decision-making purposes. While prior generations of A/B testing practitioners may have had expertise in engineering or statistics, it is clear that in industries like online retail and digital marketing, many individuals running experiments may have little formal training in these areas.⁴

Next, we reiterate that the most common paradigm used for reporting and interpreting the results of experiments in the A/B testing industry is based on simple hypothesis testing, where frequentist *p*-values—typically converted into metrics described with simplistic misnomers such as “confidence”—are used as the primary indicator of statistical evidence. In Appendix A (Figure A.2), we document how visual cues in online testing tools encourage experimenters to view tests with “confidence” above certain thresholds (e.g., 95%) as being “significant”, “conclusive”, or “actionable”; tests with “confidence” levels below the relevant thresholds are labeled with tags such as “pending”, “not enough data”, or “inconclusive”. Importantly, A/B testing dashboards are frequently updated with real-time data throughout the course of an experiment, with “confidence” levels being re-calculated after each batch update.

This combination of factors has led many practitioners and statisticians to highlight the potential for *p*-hacking (through sample size flexibility) and false discovery in A/B testing (Draper, 2016, Miller, 2010).⁵ The core statistical concern at the heart of this literature is that, to maintain their nominal false positive rates, frequentist hypothesis tests are intended to be calculated exactly once at the conclusion of a pre-specified data collection period.⁶ It has long been known analyzing simple *p*-value-based hypothesis tests at multiple points throughout the data collection process can significantly inflate an experiment’s false positive rate (Anscombe, 1954); this problem, by which an analyst wishes to evaluate the results of an ongoing statistical analysis while still con-

⁴Some commentators have stated that these platforms were specifically “designed for the uninformed” (Kohavi, 2018).

⁵It was also around this time that similar issues around multiple comparisons and false positives emerged as a prominent issue in several fields of applied academic science, including psychology, economics, and biostatistics; some commentators explicitly use the academic discussion around *p*-hacking as motivation for calling for better industrial practices (Walker, 2015).

⁶In the original circumstances in which Ronald Fisher developed his theory of *p*-values, this assumption was entirely reasonable. His analysis was primarily motivated by the problem of selecting potato varieties that best responded to fertilization, a context in which results can only be observed at a discrete point in time (after an entire season’s harvest). However, contrast this context with modern A/B testing, in which experimental results accumulate over time and can be observed continuously and instantaneously.

trolling their false positive rate, is known as “sequential testing” and dates back to Wald (1945). Indeed, practitioners in other contexts such as medicine and psychology have been aware of the “calculate once” limitation inherent to frequentist p -values for decades and developed numerous statistical techniques that allow for valid and robust sequential testing (Fiske and Jones, 1954).

Despite this long stream of literature on the problem of sequential testing and its consequences on the interpretation of classical p -values, the extent to which real-world users of A/B testing software are aware of these issues is very much unclear. There are numerous examples of analysts that have documented the ways in which they or their clients specifically stopped or extended A/B tests based on whether or not the results were “statistically significant” (Borden, 2014, Flory, 2021). There are also examples of testing platforms themselves remarking that this behavior is common among their customers (Johari et al., 2017). Further, there is some work based on data from Optimizely (with co-authors that were at the time employed by the platform) that claims to find evidence for this type of p -hacking behavior among users of that platform (Berman et al., 2018).

On the other hand, there are by now many blogs, papers, and articles discussing the issue of p -hacking and continuous monitoring in A/B testing. Further, most testing platforms do mention somewhere in their product or documentation the importance of setting an experiment’s sample sizes prior to beginning a test (though the prominence of this information varies across platforms). Given the availability of all this information, it is possible that practitioners have internalized best practices around stopping behavior and are able to avoid what was once a common pitfall. This being said, the mere fact that so many articles on A/B testing find it pertinent to mention the problem of p -hacking through sample size flexibility may itself be indicative of the prevalence of this behavior among practitioners.

2.3.2 Incentives and behavioral factors around statistical rigor in A/B testing. While prior research has demonstrated the potential for p -hacking in digital experimentation and outlined alternative methods for valid sequential testing, few researchers have explicitly addressed the role of incentives and the potential behavioral motivations of practitioners who run digital experiments. We take this opportunity to discuss several factors that we hypothesize might affect whether or not industrial practitioners p -hack. First, consider the following factors that might motivate p -hacking behavior in the context of A/B testing:

- *Significance testing is not intuitive and its goals are often misunderstood.* Given that many notions in classical frequentist statistics—significance, p -values, confidence intervals, and hypothesis testing—are not intuitive for many students (Hubbard, 2011), many researchers and practitioners misunderstand the goal of experimental analysis as achieving statistical

significance rather than recovering true model parameters (and evaluating the applicability of the model to the problem at hand) (McShane and Gal, 2016). These misconceptions about how to conduct hypothesis tests and interpret their results can result in a number of “questionable research practices” without full knowledge of how such practices undermine statistical validity (John et al., 2012). Analytic flexibility combined with a misunderstanding that induces a bias toward statistical significance—whether the bias is explicit or implicit—is well known to result in p -hacking, even among analysts with rational, well-incentivized, and benign intentions (Gelman and Loken, 2013). An analyst using A/B testing software who mistakenly believes the purpose of an experiment is to gather data until the result is statistically significant would unwittingly engage in p -hacking behavior.

- *Misaligned incentives and moral hazard within the organization.* An alternative explanation for p -hacking behavior is the presence of misaligned incentives. Work by Hall and Hasan (2020) and Ghosh et al. (2020) highlights how the adoption of A/B testing has the potential to interact adversely with the conflicting incentives across different layers of organizational structure, resulting in outcomes that can negatively affect firm performance. As for how such incentives may play into p -hacking behavior specifically, consider the role of principal-agent dynamics between managers and employees or clients and marketing agencies (Baker, 1992, Holmstrom and Milgrom, 1991).⁷ If an analyst is rewarded for finding “significant” results—but verification of these findings is inherently costly for the principal— p -hacking may result as a consequence of this incentive misalignment. Such dynamics may play out both *within* firms (e.g., between managers and their marketing teams) and *between* contracting firms (e.g., between a firm and a marketing agency).
- *Misaligned incentives between testing platforms and their customers.* As discussed earlier, experimentation platforms often treat tests that reach nominal significance thresholds with discontinuous distinction. There are, of course, benign reasons that platforms might be designed this way. Some authors have claimed that the developers of these platforms themselves are (or were at some point) misinformed about how to use p -values, and thus merely made an honest “flaw” in how they interpret them for their customers (Kohavi, 2018). Even if employees that build A/B testing tools based on frequentist statistics become aware of the problem with continuous monitoring, there are natural reasons (technical debt, having to admit prior techniques were inadequate, resistance to change by users) that may prevent them from changing their systems to reflect these concerns. It should be noted, however,

⁷Work by Spiess (2018) specifically discusses a principal-agent model of p -hacking-type behavior, though in the academic rather than industrial context.

that some analysts have suggested that experimental platforms have incentives for the continued use of design paradigms that result in an abundance of false positives (Gamber, 2019). Indeed, many marketers may be unaware of the sample sizes actually required to run truly well-powered experiments (Lewis and Rao, 2015).⁸ This may be prohibitive for many businesses of small and medium scale, a fact that experimentation platforms may not be eager to emphasize in their marketing materials or product documentation.

While the factors discussed above provide explanations for why p -hacking would be common in A/B testing, there are also very plausible reasons for why this might not be the case:

- *Statistical literacy.* Firms that use A/B testing platforms have a revealed preference for data-driven decision making and causal inference. The very act of adopting A/B testing may indicate a certain level of sophistication in understanding the value of experimentation and statistical analysis. As a result, they may be more likely to possess the knowledge and skills necessary to use these tools correctly, without resorting to p -hacking.
- *Aligned incentives.* Despite the possibilities for misaligned incentives described above, there are clearly forces working in favor of incentive alignment that might discourage p -hacking. 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.
- *Practical/logistical limitations on p -hacking.* Further, there are practical reasons why the type of p -hacking we have described—in which an analyst stops when, or waits until, a desired confidence level is achieved to stop data collection—might be rare. For one, managers may be trading off the effects of multiple metrics, without a strong preference about the significance of any particular outcome. Also, firms may have a natural cadence of meetings (say weekly, bi-weekly, or monthly) during which test results are reviewed and implementation decisions are made; this would prevent the most egregious types of truly *continuous* monitoring that are known to inflate false discovery.
- *Organizational learning.* It is possible that firms have learned to avoid p -hacking, through either the extensive literature on continuous monitoring or their own experience using ex-

⁸For a website with a baseline conversion rate of 3%, detecting a 10% lift in relative terms (using common frequentist standards of 80% power at a 5% significance level) requires over 100,000 observations.

perimentation platforms. Part of the appeal of A/B testing is that it enables managers to rigorously test their hypotheses and learn for themselves what works in their business. The instrumental purpose of an A/B test is to determine what action to take, which means there is often a very short feedback loop between an experiment and its implementation, allowing analysts to develop their own intuition about the trustworthiness of a “significant” result. Indeed, recent work by Berman and Van den Bulte (2022) suggests that A/B tests on online platforms have a relatively high false discovery rate at traditional significance thresholds (on the order of 20%-30%). Firms with experience implementing “significant” A/B test results may learn that significance levels are only a noisy signal of a test’s true effects, potentially working to discourage the targeting of classical thresholds of significance. Experts also specifically recommend running “A/A” tests (where two identical variations are tested) and repeat testing (where the same experiment is run twice) for developing intuitions about what type of statistical results can or cannot be trusted, which may further discourage the weight that firms place on “statistical significance” as a decision criteria (Kohavi et al., 2020). Thus, despite the academic preoccupation with “significance”, real-world managers have the opportunity to learn directly through experience what thresholds of evidence they require before making consequential decisions.

In sum, it appears ambiguous whether one should expect to empirically observe evidence for p -hacking of real-world A/B tests run by profit-seeking firms. Clearly, some literature suggests that p -hacking is a common practice in digital experimentation. Indeed, such a finding would be consistent with the by-now extensive literature on p -hacking in academic science, where it is known to be quite prevalent. On the other hand, there are also very plausible reasons laid out above for why this behavior might be rare in industrial settings, in which economically motivated actors face a different set of incentives from analysts in academic contexts. Given that A/B testing is a prominent example of how academic rigor and scientific reasoning are spilling over into industrial practice, understanding the extent to which the problem of p -hacking also carries over may have significant implications for both how firms view the practice and how scientists might comparatively think about the phenomenon of p -hacking in academia. This discussion motivates our main analysis in this project, in which we attempt to empirically measure the incidence of p -hacking among users of a large A/B testing platform.

3 Empirical Context & Data Description.

3.1 Description of the platform. To investigate the phenomenon of p -hacking in A/B testing, we collaborated with a private, third-party digital experience platform (subsequently

referred to as “the platform”). The platform is a large purveyor of e-commerce SaaS solutions, including web analytics, online personalization, and A/B testing. At the time of our data collection, the platform’s tracking software observed over 100 billion pageviews annually. An important difference compared to some testing platforms is that the platform studied here does not have a free tier of service; to begin using the service, a company must make contact with the platform’s sales team and set up an account. Pricing is proprietary and depends on the exact bundle of products/services a company uses, but can be expected to be in line with other enterprise B2B SaaS products. Like many other digital experience platforms, the majority of their customers use experiments for conversion rate optimization (as opposed to non-inferority type testing). Prototypical tests run by firms on the platform consist of marketing interventions (sales, promotions, product recommendations) or website design modifications (affecting button designs, page layouts, and promotional copy) designed to increase the probability of customer purchases on a firm’s website.

3.1.1 Testing Interface. Several aspects of the platform’s interface are important to discuss. Much like other third-party testing tools, firms using the platform must integrate the platform’s tracking and testing script into their website’s codebase. Once a customer’s account is created and their website is configured with the platform’s technology, the customer can use the platform’s WYSIWYG interface or custom CSS/Javascript injection to create an intervention to be tested in an experiment. Once an experimental intervention is designed by the customer, they can deploy it to their website. Throughout the course of an active A/B test, the platform’s technology allows it to manage all user randomization, analytics measurement, statistical calculations, and reporting of results.

By default, eight dependent variables are analyzed and shown to the analyst throughout the course of 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 95.6% of experiments in our sample, “conversion rate” is specified as a target metric; outside of this, the most common target metrics are session revenue and add-to-cart rate.

3.1.2 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 (or lift), standard errors, and “confidence” levels. These confidence levels are defined as 1 minus the p -value associated with a null hypothesis test for each metric. In the

case of binary outcomes, p -values are computed using the two-sided Z -test described in Section 2.2; p -values for continuous outcomes are calculated using a two-sided, two-sample t -test for equal means with a pooled variance estimate. The “confidence” associated with each outcome is only computed and displayed to the user once an experiment meets the platform’s statistical reporting criteria. For continuous metrics, the platform requires 30 observations in each treatment arm; tests on binary metrics have an additional requirement to reach a minimum of 10 observations in each of the possible outcomes. A stylized version of the dashboard which is functionally similar to the real interface, is shown in Figure 2. 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 figure.

Figure 2: Simplified recreation of test result dashboard interface

| Actionable | | Confidence | Lift | Standard Error |
|-----------------------------|-----------------|-------------------|-------------|-----------------------|
| Time on site | | 96% | 0.30% | ±0.12 |
| <hr/> | | | | |
| 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 |

Notes: This image recreates a stylized version of what analysts see when logging onto the A/B testing platform to examine the results of their experiments. All eight dashboard metrics are visible by default, with one specific metric highlighted as a “Target Metric”. If and when the “confidence” associated with a metric exceeds 95%, the metric is moved from the “Pending” section of the dashboard to the “Actionable” section.

One characteristic of 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 experimenter 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 “actionable” portion and shown in a different color. Through the use of these design cues, the platform clearly reinforces the importance of the 95% significance threshold and, in a not entirely subtle way (directly labeling significant results as “actionable”), encourages firms to react to their experiments when they reach this threshold.

3.1.3 How do firms end an experiment? When a test is running, the platform has a simple “stop” button that stops allocating a firm’s website traffic to the experiment. If a firm wants to

actually implement a variation (i.e., send all traffic to the treatment that performed better), the platform has a dedicated process for deploying non-experimental interventions.

3.2 Data collection and inclusion. We began our data collection by identifying the population of two-armed experiments (i.e., literal “A/B” tests) conducted by all US-based firms using the platform between January 2014 and February 2018. We constructed the sample for our main analysis in two stages: First, we identified each test’s target metric and retained only tests from this population that had sufficient data to meet the platform’s statistical reporting criteria. This ensures that for all the experiments in our sample, analysts would have had the opportunity to see the platform’s computed “confidence” value associated with each test’s target metric. In total, this amounted to 2,485 experiments from 242 different companies.

Upon inspection, we identified a number of experiments that appeared to both begin and end at the same time (within 15 minutes of each other). After further investigation and based on discussions with employees at the platform, we determined that it is common for firms to create multiple versions of the same test within the platform’s interface for the purposes of measuring effects on different user subpopulations. At the time of this study, the platform did not have the ability to report *post-hoc* subpopulation analysis. Thus, if an analyst wanted to measure the effect of an intervention on both mobile and non-mobile users (for example), the only way to do so would be to create two experiments in the platform’s interface, one targeted to mobile users and the other targeted to non-mobile users. While this would appear as two separate tests in our sample, for the purposes of studying experimenter starting/stopping behavior, these separate tests are better thought of as arising from a single experiment. However, this behavior creates some ambiguity in how to incorporate such tests into our sample. Assuming experimental outcomes influence the timing of firm stopping decisions, we believe that among concurrent experiments, the first to conclude is the most likely driver of this behavior. Thus, for the purposes of constructing our main study sample, we retain only the test that ended *first* among all such groups of concurrent tests; this results in a sample size of 2,270 tests from the same 242 firms. While our exposition here is focused on this particular sample, we have also conducted all of our analyses using other data-inclusion policies (retaining the whole sample of 2,485 tests; excluding all concurrent tests; retaining the test in a group of concurrent tests with the smallest p -values) and find nearly identical results to those reported here.

We now use the following sections to elaborate more on the characteristics associated with the firms in our sample and provide a number of summary statistics about the included experiments.

3.3 Description of firms in sample. Given that the platform specializes in analytics solutions for online retailers, essentially all firms in our sample are engaged in some form of direct-

to-consumer e-commerce. Within this category, the vast majority (93%) of firms are in discretionary consumer retail, most commonly apparel and home goods, with a smaller representation of firms in specialty segments (e.g., travel, electronics, cosmetics, and nutritional supplements). The remaining 7% of firms in our sample are software, telecommunications, or media/publishing companies. Though there is some heterogeneity across these industries, all the companies in our sample use their website to conduct paid transactions in exchange for goods or services. Thus, while the meaning of a “conversion” can vary dramatically across the web more generally—indicating anything from the act of clicking a link to creating an account—in our context, this word is predominantly meant to refer to the act of a customer making a monetary transaction.

To get a slightly better sense of the size of firms in our dataset, we used a third-party business intelligence service that maintains a database of organizational characteristics that can be referenced by a company’s domain name. The summary statistics for various indicators of website traffic, company age, and social media presence are provided in Table 1. There is clearly a diverse set of firms in our sample, ranging from small online businesses to large public companies with substantial offline retail operations. While a site’s “Alexa Global Rank” is based on various factors and should not be used as an exact measure of site traffic, some analyses suggest that sites near the median ranking in our dataset of 46,900 have somewhere on the order of 13,000 daily unique visitors to their website.⁹ The fact that the median founding year of a firm in our dataset is 1983 suggests that a substantial portion of our sample is comprised of legacy (i.e., not digital-native) retailers that have had to develop e-commerce capabilities in addition to their traditional offline competencies.

Table 1: Firm Characteristics

| Variable | Mean | Std. Dev. | Min. | Median | Max. | Coverage |
|---------------------|-------|-----------|------|--------|------|----------|
| Number of employees | 8,318 | 25.3K | 5 | 930 | 203K | 89% |
| Alexa Global Rank | 215K | 744K | 106 | 46.9K | 7.8M | 100% |
| Year Founded | 1967 | 42.1 | 1812 | 1983 | 2015 | 86% |
| Twitter Followers | 740K | 4.5M | 0 | 35.4K | 50M | 76% |

3.4 Experiment-level data. We now turn to the details of the experiments themselves in our main sample and begin to examine some of the factors that may help us identify p -hacking behavior. As mentioned, we have a total of 2,270 separate experiments in our sample from 242 firms; for each experiment, we observe all outcomes relevant to each treatment arm’s performance on the eight outcome metrics that are tracked by the platform. From this, we can derive the p -value that was used to calculate the “confidence” metrics that were used to mark a test result “Actionable” vs. “Pending”. On average, there are 9.4 two-armed experiments per firm, which

⁹See <https://netotraffic.com/alexa-traffic/> and <http://domain.tips/alexa-rank-traffic-converter/>

ranges from 1 to a maximum of 82; the distribution of experiment count by firm is plotted in histogram form in Figure 3. We have provided further experiment-level summary statistics in Table 2. The median experiment length in our dataset is 24.1 days and the median number of sessions per experiment is near 42,000.

Figure 3: Distribution of test count by firm

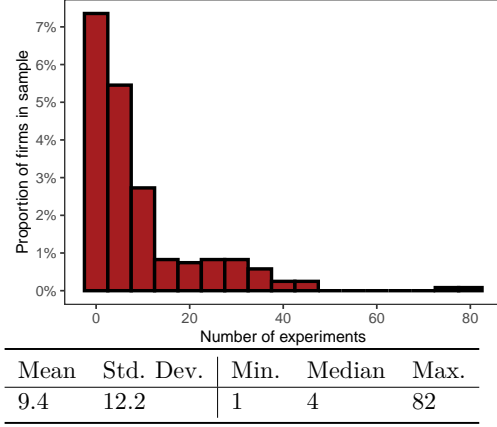
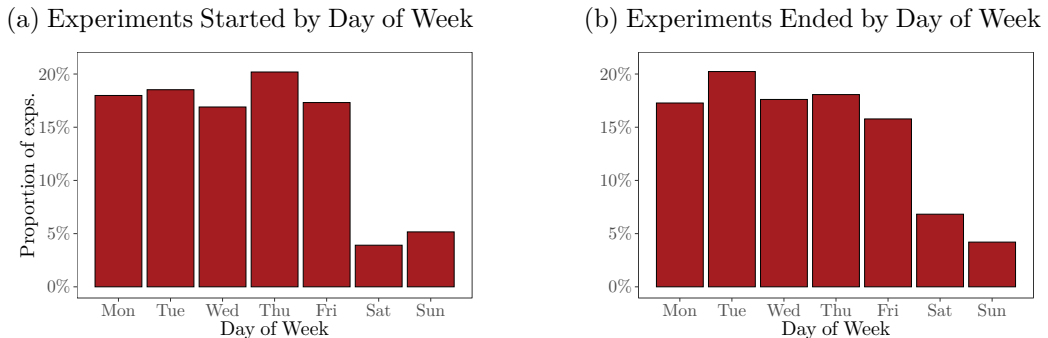


Table 2: Experiment level characteristics

| | Median | Mean | Std. Dev. |
|-----------------------------|--------|---------|-----------|
| Length of experiment (days) | 24.6 | 41.1 | 44.5 |
| Number of sessions | 41,780 | 231,127 | 440,690 |
| Conversion rate | 0.042 | 0.098 | 0.156 |
| Average order value (\$USD) | 109.79 | 186.29 | 273.81 |

3.4.1 Experiment stopping behavior. The primary question we wish to study in this project is if and how an experiment’s p -value affects a firm’s decision to end the test. To fully understand this phenomenon, it is useful to know whether other factors drive firms’ test-stopping behavior. To that end, we now present some model-free results that partially characterize when firms choose to stop their tests. First, we can see that if we plot histograms across the days of the week that experiments begin and end (Figure 4), there is a clear pattern in both graphs showing how both types of action are less common on weekends. Thus the day of the week seems to have at least some role in determining when firms end their experiments.

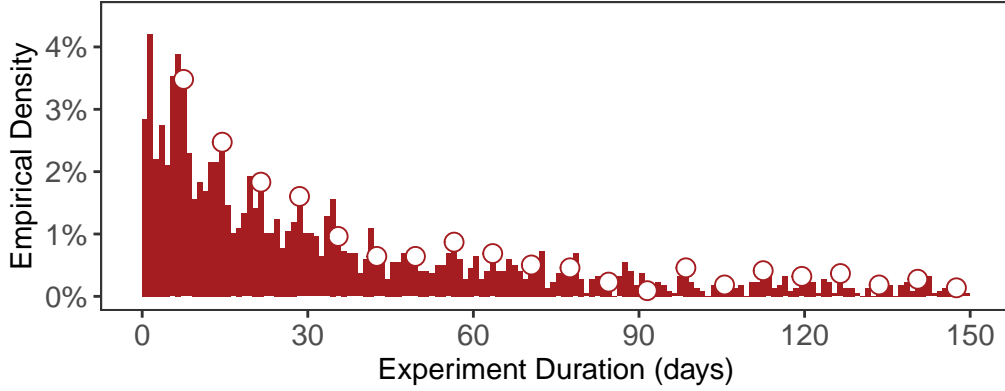
Figure 4: Starting/Stopping Behavior By Day of Week



We can see another effect at play by plotting the histogram of experiment duration (Figure 5). In this plot, we have added white circles to the tops of the histogram bars that are multiples of seven—i.e., experiments that last in increments of one week. As can be seen by the relative peaks at these values, it appears to be a common practice among firms to run their experiments for one-week intervals, with the most common experiment length being exactly one week. Again, this suggests that

logistical factors and common practices around test duration play a role in firm stopping behavior.

Figure 5: Distribution of experiment length



Notes: Histogram of experiment durations binned by day. White circles highlight histogram heights at multiples of 7 days.

4 *p*-hacking analysis.

We now turn to the question of whether the *p*-values (or, equivalently, “confidence” levels) affect the timing of when firms choose to terminate their tests and discuss ways we might detect this behavior. Prior research on meta-analytic techniques for detecting *p*-hacking in scientific literature has suggested that the type of continuous monitoring behavior alluded to in this context would result in a distribution of *p*-values with a disproportionate number of experiments just below the 0.05 cut-off; in particular, it has been suggested by Simonsohn et al. (2014) that this bunching should occur in the interval between $[0.04, 0.05)$. Thus, one way to observe *p*-hacking behavior in our dataset is to look at the empirical distribution of the *p*-values when experiments are stopped. If we assume that, for some proportion of experiments in our sample, firms were continuously monitoring their *p*-values and stopping experiments when or waiting until 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.

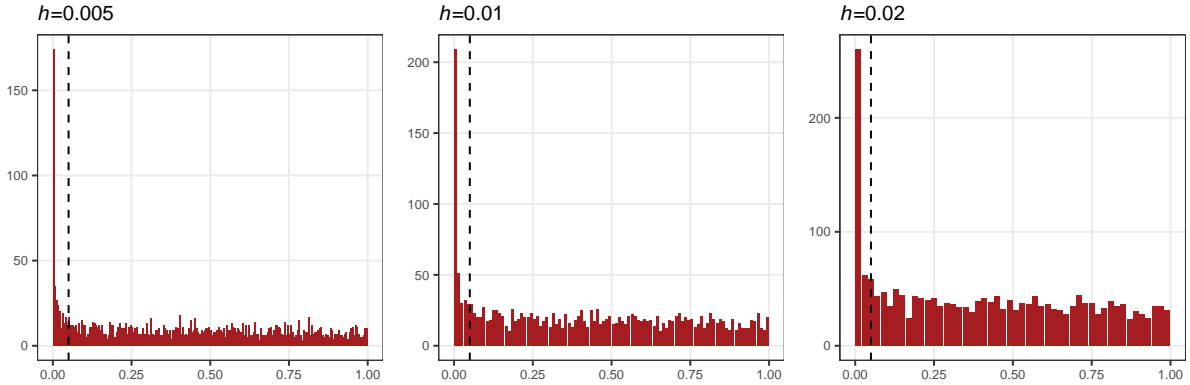
Given all the information visible to firms in the platform’s reporting dashboard, we now turn to the task of describing exactly how *p*-hacking behavior might manifest itself in our data. As a starting point, we assume that firms primarily pay attention to the “target metric” that they specified at the beginning of their experiment. In this case, we imagine that while firms can see statistics associated with other metrics, the target metric outcome plays the most significant role in driving stopping behavior. There are other ways that *p*-hacking behavior might manifest; for example, perhaps analysts monitored outcomes from all metrics, and stopped whenever *any* of their confidence levels exceed 95%. Another possibility is that firms pay attention to only a subset of metrics, such as revenue or add-to-cart rate, and discount the effects measured on time-on-site or pageviews. Because the platform gives special visual distinction to each test’s *target metric*,

we believe analyzing the distribution of p -values associated with this particular metric is the most reasonable place to look for p -hacking behavior initially. While our primary analysis is centered on the p -values associated with each test’s target metric, we explore ways of incorporating other metrics into our analysis in Section 4.5.

4.1 Model free evidence. In light of the discussion above, we focus our main analysis on the p -values associated with the target metric from each A/B test at the time it was concluded. While firms would have seen these reported as “confidence” levels, we will consider the data’s untransformed parameterization as p -values and focus on the 0.05 significance threshold. We can begin looking for evidence of p -hacking behavior by investigating the raw distribution of these p -values near this threshold.

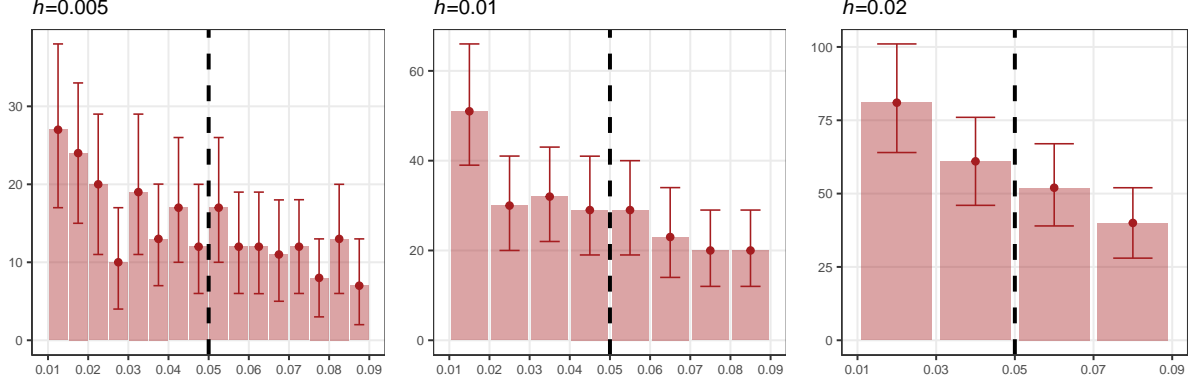
In Figure 6, we have plotted histograms of our data for three different bin widths (denoted $h = 0.005, 0.01, \& 0.02$), along with a dashed vertical line at 0.05 to facilitate inspection at this critical threshold. We have further zoomed in on the region around 0.05 in Figure 7, where we have also plotted bootstrapped 95% confidence intervals around the height of each histogram bin. Following our discussion above, a natural way to detect *discontinuous* behavior near this threshold is to compare the number of p -values in the bins on either side of the threshold. A cursory visual analysis of these plots seems to suggest that there is no evidence for an abundance of p -values just below the 0.05 threshold; the bootstrapped confidence intervals of the bins nearest the threshold seem to overlap in all cases.

Figure 6: Histograms of raw data (dashed line at $\tau = 0.05$)



We can also perform a more direct analysis of the distribution of p -values near the threshold by comparing the bin heights immediately above and below 0.05. In particular, for a given bin width h , we consider the number of experiments to the left of the threshold, denoted n_l , and the number of experiments to the right of the threshold, denoted n_r . A simple way to investigate discontinuity at the 0.05 threshold is to test against the null hypothesis that the number of observations below

Figure 7: Histograms near $\tau=0.05$ threshold, with 95% bootstrapped confidence intervals



Notes: Confidence intervals are calculated by sampling (with replacement) the entire distribution of p -values in our dataset 10,000 times and counting the number of p -values that occur in each bin region for each resampling; the bottom and top of the confidence bands correspond to the 2.5th and 97.5th percentile of bin counts across samples for each bin region.

the threshold does not exceed the number to the right of the threshold: $H_0: n_l \neq n_r$. In Table 3, we show the empirically observed values of these counts across the three different bin widths. We also report the bootstrapped p -values associated with the (two-sided) null hypothesis above, calculated by sampling (with replacement) the entire distribution of p -values in our dataset 10,000 times and calculating the proportion of samples for which the sampled value of $|n_l - n_r|$ is as large or larger than the empirically observed value of $|n_l - n_r|$. In the table, we also report the 95% confidence interval as the 0.0275 and 0.975 quantiles of the sampled distribution of the difference $n_l - n_r$.

Table 3: Bootstrap tests for discontinuity in distribution of p -values at 0.05 threshold

| Bandwidth | Obs. in [0.05-h,0.05) n_l | Obs. in [0.05,0.05+h) n_r | Empirical difference $n_l - n_r$ | 95% bootstrap confidence interval | Bootstrap test of discontinuity |
|-----------|-----------------------------------|-----------------------------------|--|---|---------------------------------------|
| $h=0.005$ | 12 | 17 | -5 | $[-16,5]$ | $p=0.54$ |
| $h=0.010$ | 29 | 29 | 0 | $[-15,15]$ | $p=0.51$ |
| $h=0.020$ | 61 | 52 | 9 | $[-12,30]$ | $p=0.52$ |

Notes: Bootstrap values are calculated by sampling (with replacement) the entire distribution of p -values in our dataset 10,000 times. We report the 0.0275 & 0.975 quantiles of the distribution of the sampled difference $n_l - n_r$ as the confidence interval. p -values are associated with the two-sided null hypothesis $H_0: n_l \neq n_r$ and are reported as the proportion of samples for which the sampled value of $|n_l - n_r|$ is as large or larger than the empirically observed value of $|n_l - n_r|$.

As can be seen in all cases reported here, the p -values all fall well above conventional significance levels. This is to say we fail to find evidence of strategic or otherwise discontinuous behavior by experimenters around the 0.05 threshold in our sample.

4.2 Other techniques for detecting p -hacking. In our research, we were able to find several methods from prior work that investigate problems similar to ours. Crucially, for the purposes of our analysis, we uncovered key shortcomings associated with both the simple bootstrap approach above and the existing methods reviewed below. We use this to motivate a simple extension of prior models, which we demonstrate has favorable properties for detecting p -hacking behavior

in our context. We briefly review prior work below, and then introduce our model and apply it to our data in an attempt to further contextualize the evidence for p -hacking behavior in our sample.

4.2.1 Generic density estimation techniques. A key stream of existing literature related to our problem comes from econometric and statistical methods for discontinuity detection in empirical data distributions. This problem of detecting discontinuities in *density functions* is a fundamentally different problem than detecting discontinuities in dependent variables, as is commonly the goal in standard regression discontinuity design. An important method for density discontinuity detection developed in recent years is that of Cattaneo et al. (2018)—denoted “**rddensity**”. This technique employs local polynomial regressions to flexibly approximate the density function on either side of the threshold of interest while taking into account many of the subtle issues that make density estimation difficult (e.g., bandwidth selection, kernel choice, bias near discrete boundaries). Because of this, **rddensity** is frequently used as a manipulation check in standard discontinuity designs and may be considered state-of-the-art for discontinuity detection in generic density functions.

4.2.2 Meta-analytic techniques. Several prior papers investigate the problem of detecting discontinuities and other abnormalities in the distribution of reported test statistics across many academic studies (Andrews and Kasy, 2019, Gerber et al., 2008). An influential paper in this stream of work is Gerber and Malhotra (2008), which introduced the “caliper test” methodology for investigating discontinuous anomalies in test statistics. This test is conducted by considering all test statistics from a body of work and selecting a small window around a relevant significance threshold. If the distribution of statistics is indeed continuous, then the number of results that fall on either side of this threshold should approximately be equal, assuming the caliper is sufficiently small. One can then perform a binomial test on whether the number of tests on the significant side of the threshold exceeds those on the non-significant side.

There are also research methods that are designed to *account for* the effects of p -hacking in the construction of some meta-analytic estimator (Simonsohn et al., 2014, Stanley and Doucouliagos, 2014, Vogel and Homberg, 2021). The methods described in these papers, however, are not well-matched to our problem for two reasons. First, most of these projects are concerned with analyzing experiments that are all designed to study a single phenomenon with a shared underlying effect size; this assumption is clearly not applicable in our context, where our experiments come from hundreds of firms testing thousands of different *ad hoc* interventions. Second, the purpose of these techniques is more to conduct meta-analytic inference that is robust to the effects of p -hacking, rather than actually detect p -hacking. While related, these techniques are ultimately

ill-suited for the problem at hand.

4.2.3 Shortcomings of existing methods. We developed a simulation procedure to compare the performance of the caliper method of Gerber et al. (2008) and the `rddensity` technique of Cattaneo et al. (2018) for detecting p -hacking in our context, and we dedicated a portion of our appendix (Appendix B.2) to studying the test performance of these methods. This analysis demonstrates the main shortcomings of these approaches, especially in relation to the approach we propose below. The primary issues are:

- *Inflated false discovery due to symmetry assumptions.* The caliper method of Gerber and Malhotra (2008) assumes that the number of observations on either side of the relevant significance threshold is *equal*. (The bootstrap method used in section 4.1 suffers from this same shortcoming.) This assumption is only theoretically true either (a) when the underlying density function is perfectly flat in the region of interest or (b) in the limit as the size of the caliper (h , the window around the relevant threshold) goes to zero, i.e., $h \rightarrow 0^+$. We show in the appendix how this can lead to an inflated Type I error (false positive) rate in the detection of p -hacking.
- *Low power (inflated false negative rate).* Generic density-based methods, such as `rddensity` from Cattaneo et al. (2018), appear to have good false positive control but are underpowered relative to other methods for detecting p -hacking reviewed in this project. Having a high-powered test is good practice in all statistical analysis, but—given that cursory analysis of our data suggests our effect might be small—this is a particularly important property to have in this project.

In a number of simulation analyses described in the appendix, we show how our approach appears to achieve a Pareto improvement in performance on desired test characteristics among these methods. In Section 4.3, we outline this approach, and in Section 4.4, we report the results from its application to our dataset.

4.3 Outline of asymmetric caliper discontinuity test. Our test aims to create a model for a distribution of p -values, assuming this distribution was generated without any discontinuous behavior near 0.05. This model helps us derive a test statistic based on the difference in the number of p -values falling on either side of the significance threshold. The bootstrap above, as well as the caliper test used in prior studies, assumed that these counts should be equal under the null hypothesis; as outlined in Appendix B.2, this assumption compromises the performance of these tests for their intended objective because it does not consider the global information about the shape of the distribution being studied. Our test corrects this shortcoming by modeling the slope of the distri-

bution near the significance threshold. This method allows for an unequal distribution of observations within a small window on either side of the threshold, making it an *asymmetric* caliper test.

4.3.1 Modeling the empirical distribution of p -values. A key requirement for our model is that we have a valid model of what a distribution of p -values should look like in the absence of any p -hacking. To do this, we draw upon existing theory and established literature in bio- and meta-statistics about how to model p -value distributions arising from many disparate analyses. Consider that for an experiment where the null hypothesis is true (i.e., there is no difference in conversion rates between treatment arms), frequentist statistical 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).¹⁰ 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 we observe in our data (Figure 7).

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 this project, we model the non-null portion of our density as a mixture of beta distributions. This approach is well-established in biostatistics and has been used extensively in many contexts to model p -value distributions (Allison et al., 2002, Parker and Rothenberg, 1988, Pounds and Morris, 2003). A full description of our model, derivation of relevant formulae, and discussion around issues such as parameter restrictions, number of mixture components in the model, and other implicit assumptions are provided in Appendix B.

In brief, we model our data as arriving from a two-stage mixture model, where 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: $\pi_1 > \dots > \pi_K$ and $\sum_{k=0}^K \pi_k = 1$. In the second stage, conditional on knowing an observation’s mixture component k_i , its value x_i will be distributed either uniformly

¹⁰While the null hypothesis may never be *exactly* true, this discussion provides an intuitive justification for the parametric form of our density function. Further, whether or not the null is precisely true in any case, we can see visually — observing the flatness of the p -value density above 0.05 in Figure 6 — that a significant proportion of the p -value data can be well-modeled with a uniform distribution.

(for $k_i=0$) or as a Beta random variable with component-specific parameters (a_{k_i}, b_{k_i}) (for $k_i>0$):

$$x_i | k_i=0 \sim \text{Unif}(0,1) \quad \text{and} \quad x_i | k_i>0, a, b \sim \text{Beta}(a_{k_i}, b_{k_i})$$

We estimate this model using the principle of maximum likelihood by maximizing the log-likelihood function, $\ell(\theta | x)$, of the observed data x over the distribution parameters $\theta = (\pi, a, b)$; this function is given by:

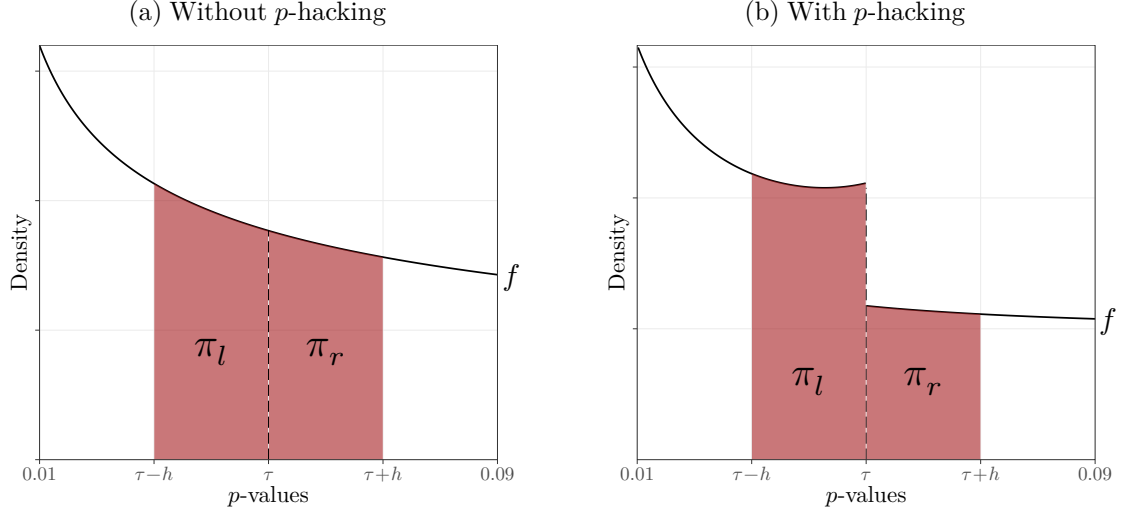
$$\ell(\theta | x) = \log \mathcal{L}(\theta | x) = \sum_{i=1}^N \log \left(\pi_0 + \sum_{k=1}^K \pi_k \left\{ \frac{1}{B(a_k, b_k)} x^{a_k} (1-x)^{b_k-1} \right\} \right) \quad (1)$$

4.3.2 Hypothesis test for discontinuity at significance threshold. The upshot of fitting the mixture model described above is that it provides a principled method for making inferences about the shape of the distribution of p -values near the significance threshold. With the information provided by the model, we can describe what the distribution of p -values *should* look on either side of the significance threshold, in the absence of any discontinuous p -hacking behavior. The visualization shown in Figure 8(a) highlights two regions within some bandwidth (or “caliper”) parameter h of the significance threshold: π_l refers to the proportion of the density function that lies in the left window and π_r refers to the proportion in the right window. These two quantities — which can be inferred as a function of the best-fit parameters of the mixture distribution described above — help us define a null distribution around our primary statistic of interest: the difference in the number of observations that fall between the left and right windows. Under the null hypothesis that our data arise from a continuous underlying density function (described by our mixture model), the counts of observations on either side of the threshold are distributed binomially as $n_l \sim \text{Binomial}(N, \pi_l)$ and $n_r \sim \text{Binomial}(N, \pi_r)$, where N is the total number experiments in our sample and n_l, n_r refer to the *count* of experiments observed in each bin. We use these distributions to calculate the null distribution of the difference between the number of observations below and above the threshold: $S := n_l - n_r$. This quantity will serve as our primary test statistic. Under the null hypothesis that our data are drawn from the fully continuous mixture model, the null distribution of S can be derived in closed form from the parameters of the mixture model (see Eq. 3 in Appendix B). In the final step of our procedure, we conduct a one-sided hypothesis test by comparing the difference between the number of observations below and above the significance threshold — which we will denote by S^* — to the estimated null distribution of \hat{S} that is derived from the mixture model.¹¹

4.3.3 Final implementation details. Two factors we have yet to discuss about our test procedure are how we decide on (a) the number of components in our mixture model and (b) what

¹¹Because an abundance of p -values below the significance threshold is the only phenomenon consistent with p -hacking, we will only accept the alternative hypothesis (that our data have been p -hacked) if the number of observations *left* of the threshold exceeds the amount on the *right*.

Figure 8: Theoretical distribution of p -values near significance threshold



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.

exactly it means to be “near” the significance threshold (i.e., how we select h). In regards to (a): For the primary results reported here, we use a model with $K = 2$ beta components (three in total, counting the uniform component). We find that among the set of integers, the choice of $K = 2$ minimizes the model’s Bayesian information criterion and maximizes cross-validated (out-of-sample) log-likelihood. However, we will report alternative analyses with other values of K after we describe our main results. As for (b): Nearly every density estimation technique requires that the researcher specify a bandwidth parameter in some form; our method is no different in this regard. As described above and visualized in Figure 8, our statistical test requires that we specify the width of the “window” around our significance threshold — given by some $h > 0$. A well-regarded technique for choosing this parameter is to use a data-driven approach such as cross-validation or machine learning (Rudemo, 1982). We will instead fix the bandwidth for our tests manually, though at several different possible values for robustness. In our context, data-adaptive approaches are prone to small-sample biases that can have unpredictable effects on statistical tests dependent on this parameter. This is explained in more detail in Appendix D, where we demonstrate how — at least for our analysis — fixed bandwidth exhibits higher power and more regularity than tests based on adaptive bandwidths.

As for how we select h , we highlight our qualitative conclusions are unchanged if we use essentially any value of h between 0.001 and 0.05. Previous researchers studying meta-analytic p -value distributions in academic research have specifically pointed to a prevalence of p -values in the 0.04-0.05 range as indicative of p -hacking behavior (Simonsohn et al., 2014). Based on this

work, our primary test of interest will use a bandwidth parameter of $h = 0.01$. However, we also describe the results of our analysis if this parameter is set to either half ($h = 0.005$) and twice ($h = 0.02$) this value; this method of reporting robustness to bandwidth choice is commonly used in regression discontinuity designs, as suggested by McCrary (2008).

4.4 Main empirical results. Having outlined the main components of our model and statistical test, we are now in a position to report the results of our discontinuity analysis. We begin by describing the fit of the mixture model; the maximum likelihood estimates for the parameters in our model are given in Table 4. Our mixture model lends itself well to graphical visualization; the MLE fit of our model on top of the empirical histogram has been plotted in Figure 9. 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 to facilitate hypothesis testing in the following sections.

Figure 9: MLE fit and empirical histogram

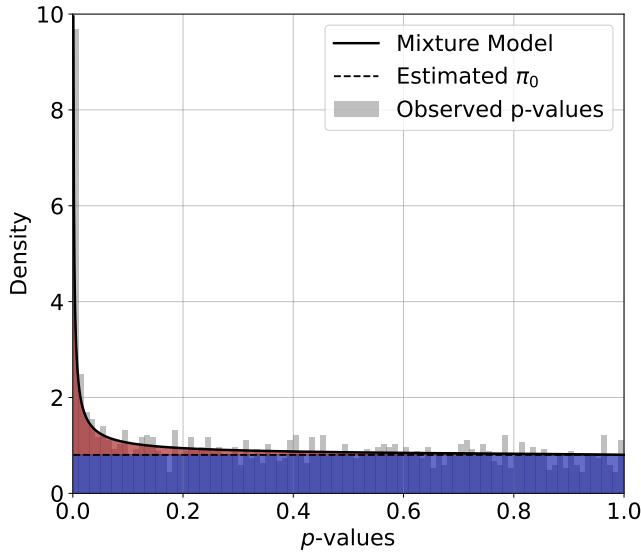


Table 4: MLE Parameter Estimates

| Parameter | Estimate |
|-----------|----------|
| π_0 | 0.802 |
| π_1 | 0.168 |
| π_2 | 0.030 |
| a_1 | 0.237 |
| a_2 | 0.010 |
| b_1 | 1.536 |
| b_2 | 1.384 |

Having computed the best-fit mixture model, we can now report the results of our hypothesis test for discontinuity in our data at the $\tau = 0.05$ threshold. These results are displayed in Figure 10, with key statistics also summarized in Table 5. Each of the three rows in the figure corresponds to a separate calculation for 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 the difference between the bin counts immediately above and below the 0.05 threshold; these have been

colored in solid red to highlight the portion of the data that is used for the local discontinuity tests. Additionally, the MLE fit of the beta-uniform mixture model is 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 for each test are displayed above the graphs of the test statistic distributions. Because we have conducted a one-sided hypothesis test, p -values are computed by calculating the proportion of the test statistic’s null distribution that lies above the observed value; this integration has been represented visually by shading the region of the null distribution above the observed test statistic.

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 ($p=0.83, 0.55, 0.37$). 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.

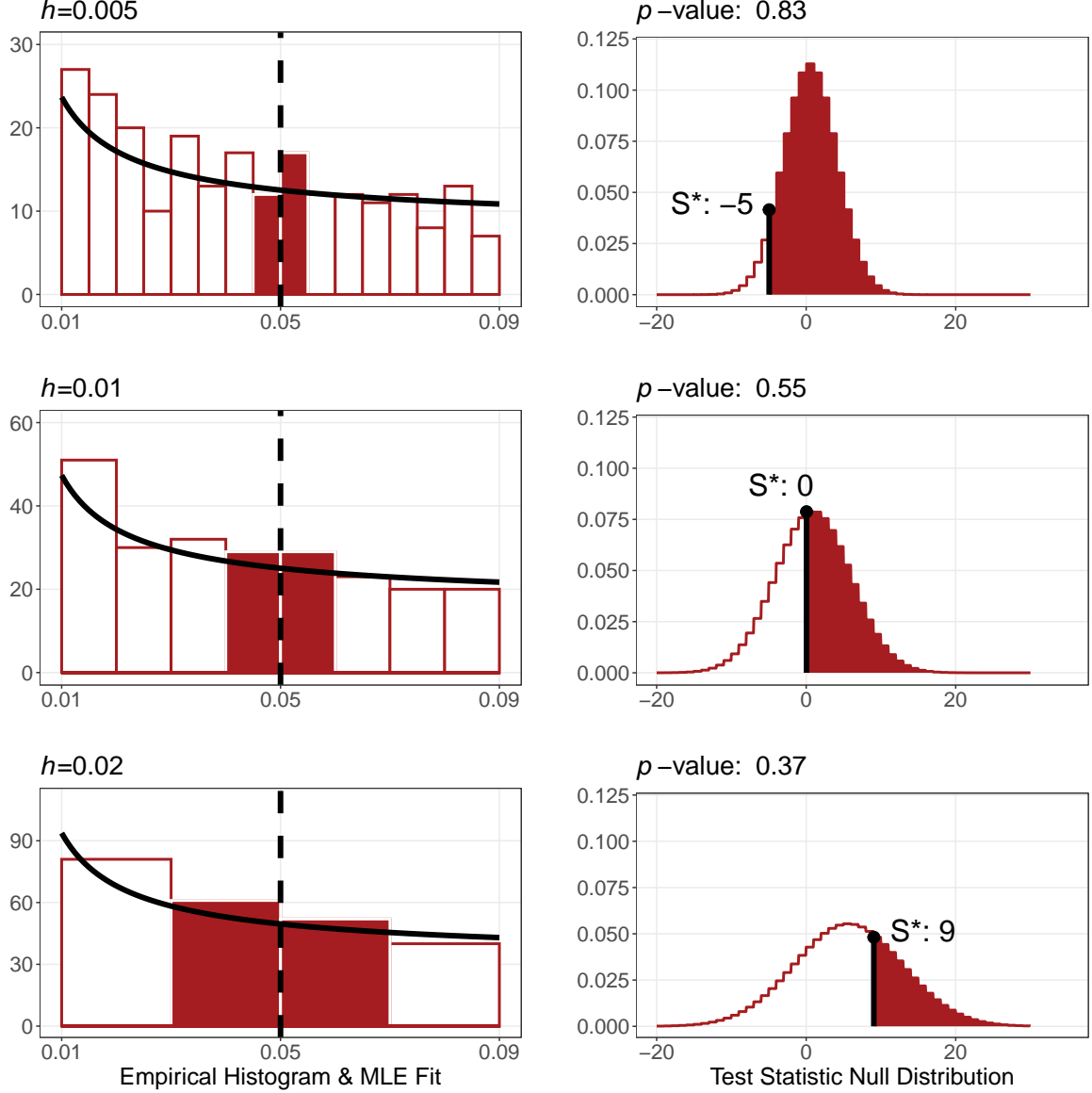
Recall that the results reported above correspond to the version of our model with $K = 2$ beta-distributed mixture components and one uniform component. To ensure our results are not sensitive to this assumption, we also performed our analysis for values of $K \in \{3, 4, 5\}$ and we report the corresponding results of our discontinuity test in Table 6. The results from these tests are all very consistent with our primary analysis, demonstrating that our null finding above is robust to this particular parameter.

4.4.1 Interpreting the null result & counterfactual 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. However, there are several steps we can take to better understand and contextualize what our results can and cannot say about the presence of p -hacking in our sample. We begin by considering the confidence intervals associated with our main effect, move on to discuss some of the nuances associated with quantifying the effect we are

Table 5: Primary tests for discontinuity at $\tau=0.05$ threshold

| | Values left of threshold | Values right of threshold | Empirical difference | Expected difference | Asymmetric caliper |
|-----------|-----------------------------|------------------------------|-------------------------|------------------------|-----------------------|
| Bandwidth | n_l | n_r | S^* | $\mathbf{E}[\hat{S}]$ | p -value |
| $h=0.005$ | 12 | 17 | -5 | 0 | 0.83 |
| $h=0.01$ | 29 | 29 | 0 | 2 | 0.55 |
| $h=0.02$ | 61 | 52 | 9 | 6 | 0.37 |

Figure 10: Results of asymmetric caliper test for discontinuity at $\tau = 0.05$



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 are shown in the left column (white/red bars), along with the MLE-fit beta-uniform mixture model (black line). In the right column, for each h , we plot the corresponding null distribution of our test statistic (derived in Appendix B, Eq. 3) and the empirically observed difference in bin heights, S^* . Taking the proportion of the null distribution that lies above S^* gives us the one-sided p -value for our discontinuity test (see Eq. 8 in Appendix B for precise definition). No p -value among the outcomes we observe ($p = 0.83, 0.55, 0.37$) approaches conventional levels of significance, indicating we cannot reject a null hypothesis that assumes continuity in the underlying density of p -values in our dataset.

trying to detect, and then describe our attempt at measuring the power of our discontinuity test.

We describe in Appendix B how one can adapt our test procedure to define both the one-sided and two-sided 95% confidence intervals around our main test statistic; these intervals are shown in Table 7, which have been calculated using the same three bandwidth thresholds as above. For now, consider the test with $h = 0.02$, which has the widest confidence intervals among our three analyses. The two-sided 95% confidence interval for the size of the discontinuity at the $p = 0.05$

Table 6: Discontinuity test results (p -values) for different numbers of mixture components (K)

| | $K=2$ | $K=3$ | $K=4$ | $K=5$ |
|-----------|-------|-------|-------|-------|
| $h=0.005$ | 0.834 | 0.828 | 0.837 | 0.841 |
| $h=0.01$ | 0.553 | 0.562 | 0.574 | 0.587 |
| $h=0.02$ | 0.368 | 0.402 | 0.426 | 0.461 |

Notes: This table reports one-sided p -values from the asymmetric caliper test, applied to our data at the $\tau = 0.05$ threshold. The twelve specifications were generated by varying the number of mixture components (K) and bandwidth parameters (h).

threshold is $(-16, 22)$. To interpret these numbers, consider that our test statistic is constructed as the *difference* between the number of p -values below and above the significance threshold. If one imagines simplistically that a p -hacked test is one that moves a result that would have fallen in the window directly above the 0.05 threshold to the window below the threshold, this would cause the difference in our test statistic to increase by two. This suggests it may be better to divide this statistic in half when attempting to translate the result into a “number of experiments that were p -hacked”. With this particular interpretation, our data would indicate no more than 11 experiments (the upper end of the 95% confidence interval divided by two) or — 0.53% of the tests in our total sample — may have been p -hacked. While such reasoning can be a useful thought experiment, we argue one should *not* interpret our results using this simple logic. This is because it is simply not true that the difference between a “ p -hacked” and “not p -hacked” experiment is its p -value would move from just above 0.05 to just below 0.05. It is instructive to understand why this is the case.

Table 7: Confidence Intervals

| | Bandwidth h | | |
|-----------------------------------|----------------|-----------------|-----------------|
| | 0.005 | 0.01 | 0.02 |
| One-sided 95% confidence interval | $(-\infty, 3)$ | $(-\infty, 10)$ | $(-\infty, 19)$ |
| Two-sided 95% confidence interval | $(-14, 4)$ | $(-14, 12)$ | $(-16, 22)$ |

In our context, p -hacking can manifest in two ways. First, tests with genuine differences between treatment arms might naturally yield p -values close to zero. If p -hacked, these tests could prematurely stop once the p -value is below 0.05, shifting it from nearly zero to just under this threshold. Conversely, p -hacking can be applied to tests with effect sizes that are effectively null (implying their p -values are uniformly distributed). In these scenarios, an experimenter might extend their data collection until the p -value happens to fall below 0.05 by chance. Without ground-truth knowledge of the actual effect sizes, distinguishing p -hacking types and interpreting a discontinuity’s magnitude is challenging. All p -hacking creates a discontinuity at the critical threshold, but without knowing true effects, we are unable to deduce the exact magnitude of the “number of p -hacked experiments”.

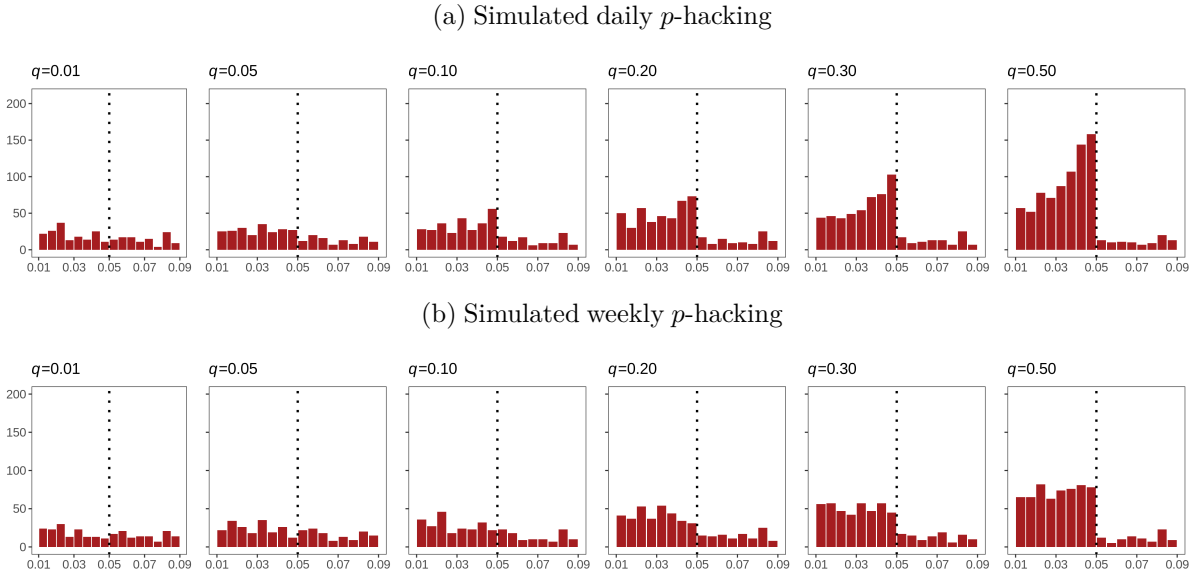
Nonetheless, it would be very useful—both for conceptual reasons and to interpret the results of our analysis—if we could quantify the “effect size of p -hacking” in some way. We propose one way to think about this effect, which is as *the proportion of experiments that were intended to be p -hacked*. For every A/B test a firm could run, we imagine two possible scenarios: In the first scenario, the experiment is run in a “natural” state without any interference of p -hacking behavior (i.e., we assume the experiment ends for exogenous reasons unrelated to the test’s p -value). In the second scenario, the test is run with the intention of being p -hacked, where the stopping time of the experiment is determined by how its p -value evolves over time. In this case, p -hacking would proceed along the lines described in the previous paragraph, where an analyst might stop a test before its “natural” lifetime, or extend it beyond its “natural” lifetime depending on the underlying data generating process.

Based on this understanding, we have developed a procedure that allows us to counterfactually p -hack the experiments in our dataset. In addition to observing the terminal p -value associated with each test in our sample, we also observe (a) the raw outcomes of the data used to derive the p -value (e.g., the number of user sessions and conversions observed in each test arm), and (b) the duration of each experiment in days. Using these data, we can simulate what the contemporaneous p -value would be throughout the duration of each test. We can then simulate what the terminal p -value of a test would be if it were p -hacked. Our full simulation procedure is described in Appendix C, but in short our process is as follows: Assume some “effect size” $q \in [0,1]$ is given (i.e., the proportion of experiments that were intended to be p -hacked); take a random sample of $\lceil qN \rceil$ experiments from our dataset and “counterfactually” p -hack them; this involves randomly dividing up the number of observations across the lifetime of an experiment, calculating the p -value at each interval of time throughout the life of an experiment, and ending an experiment early if its p -value becomes significant within its observed duration. If a counterfactually p -hacked test has not reached significance by the end of its natural duration, we simulate the effect of extending data collection for a period of time and stop the experiment at the first point a p -value dips below 0.05.

An important factor that determines precisely how “the proportion of experiments intended to be p -hacked” maps onto the shape of the final distribution of terminal p -values is the *frequency* at which each test’s p -value is checked. Truly continuous monitoring would consist of checking the p -value of a test after every new observation. That being said, this form of p -hacking would be unrealistic in our context. Based on our conversations with employees at the platform, it is more common for experimenters to check their results daily, weekly, or with a frequency somewhere in

between.¹² To visualize how these different forms of p -hacking would manifest in the distribution of terminal p -values, we have performed the simulation described above for varying proportions of p -hacked experiments $q \in [0,1]$ and plotted the distribution of resulting p -values near the 0.05 significance threshold. In Figure 11(a), we used a procedure that simulates the effect of checking p -values every 24 hours; in Figure 11(b), we perform the same analysis, except simulate the effect of checking p -values on a *weekly* basis. What is apparent in these visualizations is that checking p -values less frequently makes it more difficult to observe the discontinuity in the p -value distribution near 0.05. That being said, even if only a fraction of experiments are p -hacked, the effect of this behavior should be noticeable.

Figure 11: p -value distributions from counterfactually p -hacked A/B tests



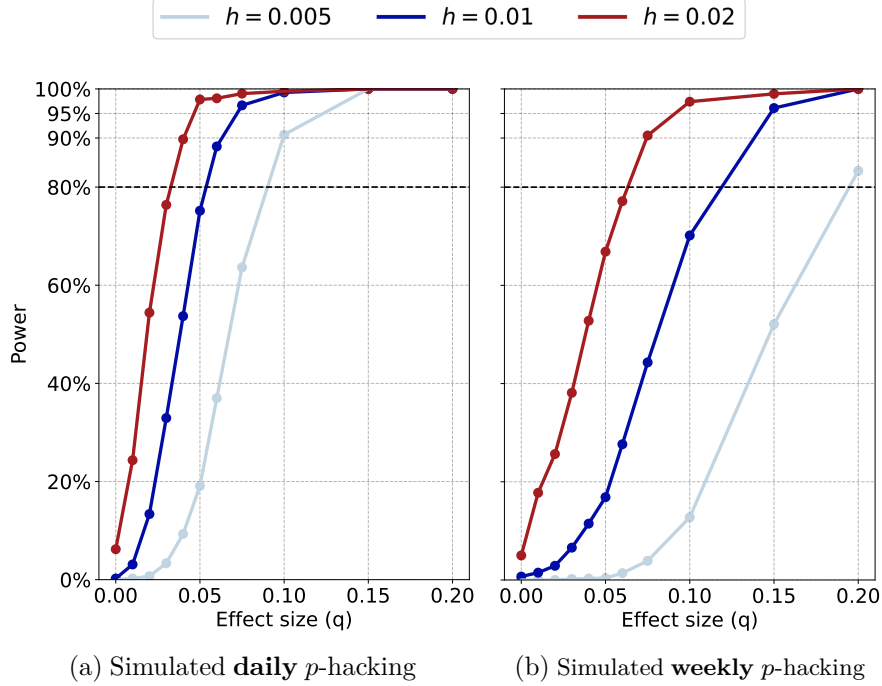
Notes: For each subplot, we randomly select a fraction q of tests in our sample and simulate what their p -values would have been throughout the lifetime of each test. In Fig. (a), experiments are counterfactually p -hacked by checking p -values once every 24 hours; in Fig. (b), the experiments are p -hacked by checking p -values once every 7 days. For full details on the simulation procedure, see Appendix C.

In fact, we can be more precise about this observation and use our simulation procedure to quantify how often our statistical test is able to detect a discontinuity at the p -hacked significance threshold. To this end, we estimated the power curve of our statistical test in the following way: For each of several values of $q \in [0,0.20]$, we took a bootstrap sample of tests from our dataset (using $N = 2,270$ and sampling with replacement), simulated the p -value distribution we would observe if a fraction q of the tests were deterministically p -hacked (i.e., stopped on the first time a p -value below 0.05 is observed), performed our asymmetric caliper test, and took note of whether our test was able to detect a discontinuity at the $\alpha = 0.05$ significance level. For each value of q , for various values of the bandwidth parameter h , and for both daily and weekly forms of p -hacking,

¹²Also recall from the discussion in Section 3.4.1 that a subset of analysts appears to make stopping decisions on a weekly basis.

we repeated this procedure 1,000 times and took note of the percentage of the 1,000 repeats in which we detected a discontinuity. (This procedure is described in detail in Appendix C.) The results of this analysis are summarized in Figure 12.

Figure 12: Power (sensitivity) analysis of asymmetric caliper method for detecting p -hacking



Notes: These are plots generated by using Monte Carlo simulations, where for each of 1,000 bootstrap samples per effect size q , we simulate the daily (weekly) time series of p -values for all experiments in our data and counterfactually p -hack a proportion q of them. For each simulation, we perform our asymmetric caliper test for discontinuity and report the “power” of our test as the percentage of times our test detected a significant effect at the 5% level. Results are grouped by bandwidth size h .

For the most sensitive version of our test procedure (i.e., with $h = 0.02$), this analysis estimates we would have 80% power to detect an effect size of $q \approx 0.035$ for daily p -hacking and $q \approx 0.065$ for weekly p -hacking. To be sure we interpret these results correctly, note that real p -hacking behavior would obviously be less deterministic than the behavior we are simulating. However, this method does capture many of the essential dynamics of real p -hacking behavior and is able to quantify the power of our testing procedure while also keeping the problem tractable and interpretable. 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. If, on the other hand, we believe p -hacking behavior occurs less frequently than once a week, then the power of our test is lower than what we report in Figure 12b. However, we believe the most common forms of real-world p -hacking behavior would very likely fall somewhere between these two extremes.

So while we cannot conclusively rule out the presence of p -hacking in our sample, we can say with reasonable confidence that if there were a positive effect of p -hacking in our dataset, it would

affect fewer than 3%–6% of the experiments in our sample. Even in the worst-case scenario that there is an effect and its magnitude is on the upper end of this range, this estimate is smaller by nearly an order of magnitude when compared to existing estimates of the prevalence of p -hacking in similar contexts (Berman et al., 2018). To summarize: we find no evidence for the hypothesis that e-commerce firms p -hack at the 5% significance threshold. Further, if even a small fraction of experimenters were systematically engaging in this behavior, the analyses laid out here demonstrate that our methodology would have the ability to detect this effect. We emphasize that across the many specifications we have analyzed, we find no positive evidence for the presence of a discontinuity in our empirically observed distribution of p -values.

4.5 Investigating other forms of p -hacking. Given that our primary analysis resulted in a null outcome, it is natural to ask if there is evidence for other forms of p -hacking behavior in our sample.

4.5.1 Alternative thresholds. One possibility we have yet to consider is that firms p -hack their A/B tests at thresholds other than the platform default of “95% confidence”. Given that the platform in our context had a very clear default around the 95% threshold, we believe it to be unlikely to find evidence for discontinuous behavior at these other thresholds, but nonetheless report the results of analyses that investigate behavior around the 90% and 99% thresholds (or the 0.10 and 0.01 thresholds in the observed distributions of p -values). Fortunately, the method we developed can easily be adapted to test for discontinuities at these other thresholds; the results of these statistical tests—conducted for varying bandwidth levels—are shown in Tables 8 and 9. We have also plotted histograms of our data near these thresholds (along with bootstrapped confidence intervals) in Figures 13 and 14. In none of these analyses do we find evidence for discontinuous behavior around the 0.10 or 0.01 significance thresholds.

Table 8: Test for p -hacking at $\tau=0.01$ level

| h | n_l | n_r | S^* | $\mathbf{E}[\hat{S}]$ | p -value |
|-----------|-------|-------|-------|-----------------------|------------|
| $h=0.001$ | 7 | 5 | 2 | 0 | 0.67 |
| $h=0.002$ | 11 | 11 | 0 | 1 | 0.37 |
| $h=0.005$ | 35 | 27 | 8 | 7 | 0.54 |

Table 9: Test for p -hacking at $\tau=0.10$ level

| h | n_l | n_r | S^* | $\mathbf{E}[\hat{S}]$ | p -value |
|----------|-------|-------|-------|-----------------------|------------|
| $h=0.01$ | 27 | 17 | 10 | 0 | 0.92 |
| $h=0.02$ | 47 | 35 | 12 | 2 | 0.86 |
| $h=0.05$ | 119 | 108 | 11 | 12 | 0.46 |

Another possibility worth considering is that firms may be using different thresholds to p -hack their experiments (e.g., one firm may p -hack at 0.05 and another at 0.10). We explore this topic in more depth in Appendix E, where we simulate more complex firm behavior and estimate the robustness of our methods for detecting p -hacking. While this type of heterogeneity would potentially diminish the power of our test, we conclude that (if a meaningful amount p -hacking were indeed occurring) we should still be able to see evidence of this behavior in the distribution of

Figure 13: Histograms of p -values near $\tau = 0.01$ threshold

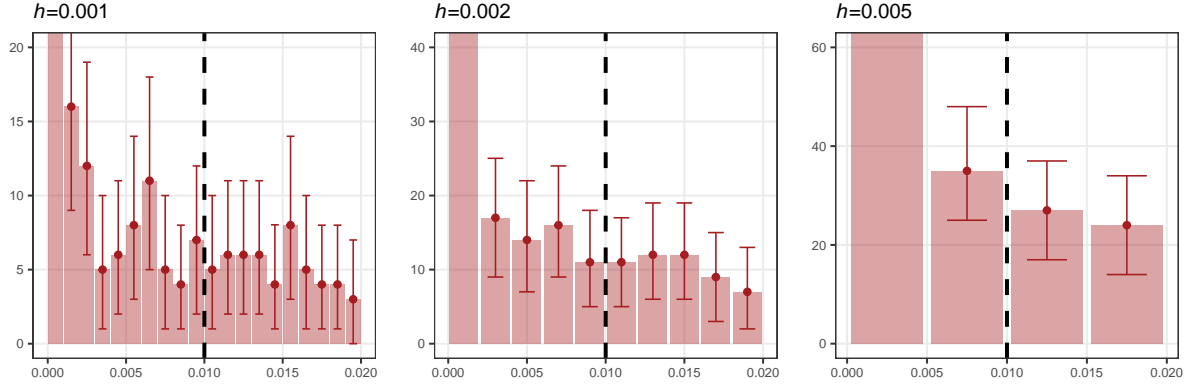
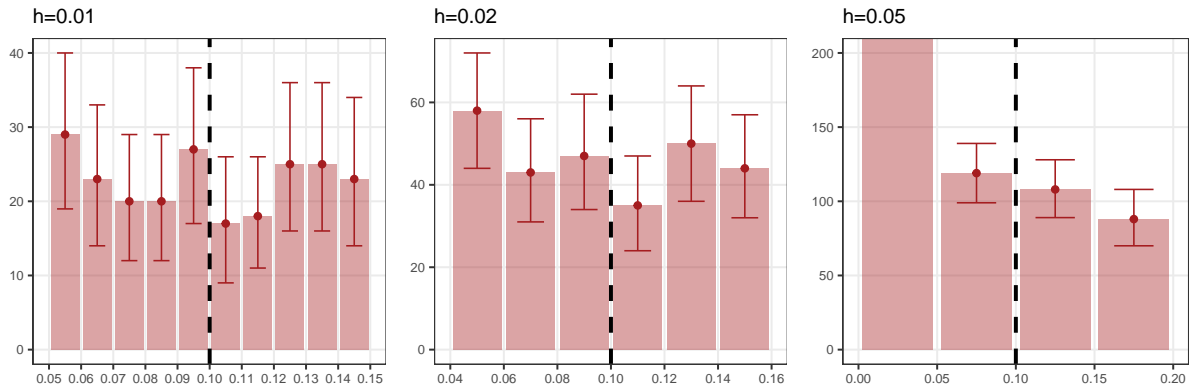


Figure 14: Histograms of p -values near $\tau = 0.10$ threshold



p -values and detect some form of discontinuity with our proposed methods.

4.5.2 Alternative (non-target) metrics on the testing dashboard. Recall (from Sec. 3.1.1) that the testing platform allows firms to choose a primary “target” metric for each A/B test; in 95.6% of tests, this metric was set to user conversion rate. But also recall that the platform reports the effect sizes and confidence levels for eight metrics by default (see Figure 2). While we have argued that the most plausible place to look for p -hacking behavior is in the distribution of p -values for the target outcome metric, there are potentially many other ways that the entirety of data reported by the testing platform could have influenced experimenters’ behavior. For example, perhaps analysts were focusing on multiple metrics of interest without going to the trouble of specifying them as “target” metrics in the platform’s testing interface. It seems reasonable—especially in our sample of firms that are all engaged in some form of e-commerce retailing—that metrics such as the add-to-cart rate and average revenue could have particular relevance.

For each outcome variable, we filter out any experiments that failed to meet the platform’s statistical reporting thresholds; in total, we are left with 16,093 p -values across the eight metrics. We plot the local distribution of terminal p -values from all eight metrics and report the results of our asymmetric caliper test applied to each of these outcomes at the $\tau = 0.05$ significance

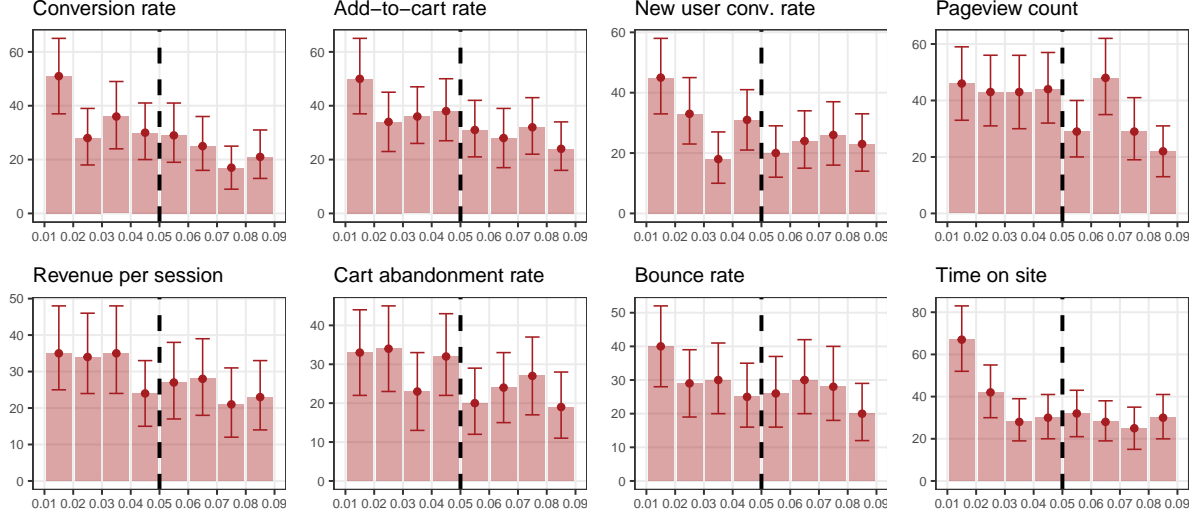
threshold. These data are shown visually in Figure 15, and the results of our test for discontinuity are summarized in Table 10. As can be seen, there is little evidence for p -hacking at the $\tau = 0.05$ (95% confidence) threshold for any of the outcome metrics visible to firms on the platform’s testing dashboard. For the “cart abandonment” and “pageviews” metrics, we do observe a p -value below 0.05 when testing for a discontinuity using the bandwidth $h = 0.01$; however, evidence for these discontinuities does not remain robust to other choices of the bandwidth parameter. Given that we conducted 24 different tests for discontinuity in this process (eight metrics, across three bandwidth values), it is not terribly unexpected that at least one of our test results appears “significant”. Also, we note in situations where we know the experiments have been p -hacked (i.e., from simulation evidence reviewed in Figure 12), we can see that test sensitivity increases with bandwidth size. However, as can be seen in the third column of Table 10, evidence for discontinuities in the p -value distributions of these two metrics using the $h = 0.02$ bandwidth is conspicuously absent. As such, we consider the total evidence for discontinuities around the significance threshold for any metric to be weak. On the whole, there appears to be little to no evidence for p -hacking behavior across different experimental outcomes.

4.5.3 Potential for heterogeneous behavior in sub-samples. Given the null results in our main analysis, a natural question to ask is if there is a sub-sample of our data that can be characterized by some dimension of heterogeneity (e.g., small firms vs. large firms, or tests with large effects vs. small effects) in which p -hacking is more likely to occur. We do not categorically rule out this possibility, but it is worth highlighting why this would be surprising to find in light of the evidence already presented. Note our analysis is specifically designed to detect the presence/absence of p -hacking in our sample. This scenario differs from other types of analysis in which heterogeneous effects across subgroups might cancel out in aggregate, resulting in a null main effect. If p -hacking occurred in any subset of our experiments, we would expect some evidence of a discontinuity to be preserved in the full sample. Though combining p -hacked and non- p -hacked p -values may introduce some noise, it should not eliminate the discontinuity entirely. So while we are not claiming that absolutely no p -hacking occurred on the platform in question, the fact that we failed to find evidence for this behavior in our main analysis suggests it must have been relatively rare across the board. In particular, recall from our analysis in Section 4.4.1 that our methods should be able to detect p -hacking if it occurred in approximately 3%-6% of the tests in our sample.

5 Discussion.

As academic and industrial sciences continue to cross-pollinate ideas and methodologies, it is increasingly important to understand when and how statistical tools adapted from one domain en-

Figure 15: Local histograms for all outcome metrics



Notes: Histogram plots of raw data near the 0.05 significance threshold, shown for each of the eight outcome metrics visible in the testing platform’s dashboard. We have also plotted 95% bootstrap confidence intervals around the observed bin counts.

Table 10: Results of tests for discontinuity at $\tau=0.05$ threshold, for all eight outcome metrics

| Metric | Bandwidth: $h=0.005$ | | | | | Bandwidth: $h=0.01$ | | | | | Bandwidth: $h=0.02$ | | | | |
|-----------------------|----------------------|-------|-------|--------------|------------|---------------------|-------|-------|--------------|------------|---------------------|-------|-------|--------------|------------|
| | n_l | n_r | S^* | $E[\hat{S}]$ | p -value | n_l | n_r | S^* | $E[\hat{S}]$ | p -value | n_l | n_r | S^* | $E[\hat{S}]$ | p -value |
| Conversion rate | 13 | 17 | -4 | 0 | 0.776 | 30 | 29 | 1 | 1 | 0.500 | 66 | 54 | 12 | 6 | 0.254 |
| Revenue per session | 10 | 9 | 1 | 0 | 0.404 | 24 | 27 | -3 | 1 | 0.694 | 59 | 55 | 4 | 6 | 0.541 |
| Add-to-cart rate | 23 | 17 | 6 | 0 | 0.120 | 38 | 31 | 7 | 2 | 0.225 | 74 | 59 | 15 | 8 | 0.238 |
| Cart abandonment rate | 15 | 10 | 5 | 0 | 0.139 | 32 | 20 | 12 | 1 | 0.043* | 55 | 44 | 11 | 5 | 0.245 |
| New user conv. rate | 14 | 10 | 4 | 0 | 0.192 | 31 | 20 | 11 | 1 | 0.065 | 49 | 44 | 5 | 5 | 0.479 |
| Bounce rate | 16 | 12 | 4 | 0 | 0.188 | 25 | 26 | -1 | 1 | 0.599 | 55 | 56 | -1 | 5 | 0.729 |
| Pageview count | 20 | 19 | 1 | 1 | 0.430 | 44 | 29 | 15 | 3 | 0.048* | 87 | 77 | 10 | 11 | 0.517 |
| Time on site | 15 | 17 | -2 | 1 | 0.657 | 30 | 32 | -2 | 3 | 0.698 | 58 | 60 | -2 | 11 | 0.858 |

hance or impair decision-making in the other. Over the past decade, firms in the e-commerce, media, and software industries have quickly adopted the use of randomized experiments—a mainstay tool of academic and medical sciences—for informing strategic and marketing decisions. Over the same time period, the potential for abusing scientific techniques—and a particular form of abuse called *p*-hacking—has attracted a significant amount of interest from practitioners, academics, and scientific commentators. Existing literature on the subject of *p*-hacking suggests it is ubiquitous in academic science, and several researchers have claimed the issue is similarly widespread in industrial contexts. In this project, we had the opportunity to investigate the behavior of 242 firms using experimentation software in an economically consequential context. We set out to determine if a certain form of *p*-hacking—in which analysts use sample size flexibility to obtain statistically significant results—is present in our sample of nearly 2300 e-commerce A/B tests. We posited (and also demonstrated) that if firms have a salient significance threshold they are trying to reach, this type of *p*-hacking would result in a discontinuity in the distribution of *p*-values. We looked for evidence of such a discontinuity across multiple significance thresholds and across multiple metrics

available to users of the A/B testing software and found null evidence in essentially all cases.

5.1 Generalization to other A/B testing contexts. As in all empirical analyses, care should be taken when considering how results from one domain may generalize to others. Though we have hundreds of firms and thousands of tests in our dataset, our research context is limited to the behavior of analysts on one commercial experimentation platform, and there may be idiosyncrasies of this platform that are important for contextualizing our results. For example, recall that the most commonly-selected target metric for tests in our sample is “conversion rate” (i.e., purchase incidence)—an outcome which may be less easy to manipulate compared to other outcomes. If firms are more interested in affecting outcomes such as “engagement”, “dwell-time”, or “click-through-rate”, interventions on these outcomes may have larger effect sizes and be more susceptible to p -hacking. Also recall our sample of firms consists mostly of well-established e-commerce businesses that may have more web traffic (and larger sample sizes) than other contexts where A/B testing is used (e.g., email marketing and web design for small businesses). On average, p -values from tests with small sample sizes are more variable and likely to change in response to new data, suggesting again that the incidence of p -hacking may be different in other settings. We also reiterate that all the firms in our sample paid a non-trivial monthly fee to use the experimentation platform in question; this indicates the firms have more than a casual interest in data-driven decision making and A/B testing, and suggests our results may not directly generalize to free experimentation tools. Further, our study will have little to say about experimentation practices at companies who use their own bespoke A/B testing tools (including prominent technology companies such as Google, Microsoft, Amazon, etc.), since the expertise, protocols, and objectives of analysts at these companies should be expected to be different from those in our sample.

This being said, we believe the characteristics of the platform we study here are such that our findings will have some degree of broader relevance. Our sample is comprised of hundreds of e-commerce firms that use A/B testing to inform consequential decisions about their websites’ designs. And we emphasize the similarities between the platform studied here and the other A/B testing products reviewed in Section 2 — in terms of their use of WYSIWYG editors, continuous reporting of experimental data via dashboard interface, and design cues applied to statistically significant test results. To be sure, future research on A/B testing in other industrial environments would be useful for helping us understand if and how different contextual factors affect p -hacking behavior. Despite the widespread commentary on the subject reviewed in Section 2.3, there exists relatively little published empirical evidence on p -hacking in commercial settings and we hope this project serves as an impetus for more inquiry.

5.2 Contextualization in light of existing p -hacking literature. There exist numerous studies on p -hacking in academic settings where it has been found to be prevalent. A natural question to consider given the results of our analysis is why a phenomenon that is widespread and easy to detect in these settings appears absent in our context. In Section 2.3.2, we discussed several explanations that are useful to review here: *statistical literacy*, *logistical limitations of p -hacking*, *aligned economic incentives between analysts and firms*, and *organizational learning*. While we cannot definitively say any particular factor drives the results in our setting, this project has provided some context for understanding the incidence of p -hacking on A/B testing platforms. To facilitate this discussion, we reiterate some of the key characteristics of our sample (described in Section 3): the users of the testing platform are mostly marketers, analysts, and web developers working for e-commerce retailers. We believe it is safe to assume the typical users of the testing platform would have had less statistical training than the typical academic researcher. Indeed, we described in Section 2 how most A/B testing platforms are explicitly designed to facilitate the use of experimental methods among marketers and analysts *without* statistical expertise. Given that p -hacking is known to occur in academic settings where analysts presumably have more statistical education, this suggests that differences in statistical literacy or formal training are not strong candidates for explaining why p -hacking appears in some contexts but not others. Further, recall from our discussion in Section 4.4.1 how we demonstrated that even if a small percentage of experiments (on the order of 3%-6%) were regularly monitored for significant statistics, this behavior would result in a reliably detectable discontinuity in our dataset’s distribution of p -values. This suggests that the type of p -hacking investigated in this project is logistically feasible as a matter of principle, and not obviously ruled out by contextual factors such as the typical sample or effect sizes of experiments on the platform.

While more research is clearly necessary, we believe the results of this project highlight the potential role of the other factors we have discussed: economic incentives and the ability of analysts to learn how to interpret statistical evidence through experience and implementation. We have mentioned the work of Berman et al. (2018), which finds some evidence of p -hacking behavior on a similar but different experimentation platform (Optimizely). While there are several differences between our two contexts, one factor we highlight is that Optimizely had a free-trial and freemium pricing strategy during their study period. 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 the Optimizely sample come from free users, this distinction between our samples may correlate with other firm

characteristics that predict p -hacking behavior. We highlight one possibility, which is that firms that have paid for a testing service are more incentivized to use it in a way that yields more robust results. To the extent that one believes employee and firm incentives are aligned, our results are consistent with the theory that industrial practitioners face at least some economic incentive to avoid engaging in p -hacking behavior.

To further develop this line of thinking, it is instructive to compare the explanations proffered above across academic and industrial contexts. In academic science, there is a putative incentive for researchers to publish true and robust results; that being said, some research has suggested the incentives for novelty and publication are stronger and not necessarily aligned with incentives for truth-finding and robustness (Open Science Collaboration, 2015, Shaw and Nave, 2023). In Brodeur et al. (2023a), the authors find evidence that academic papers in economics are p -hacked prior to formal journal submission and that reviewer evaluations are correlated with a submission’s statistical significance. The paper also documents a widespread belief among practitioners that editors and reviewers have strong preferences for “significant” results, potentially explaining the prevalence of p -hacking in settings where academic publication is a key objective of a statistical analysis.

In contrast, industrial analysts use statistical tests to make instrumental decisions in their businesses. In this setting, there is a direct connection between statistical rigor and economic value. Further, whereas in academic science, p -hacked results might only be revealed in the event of a replication or detailed analysis by a motivated methodological researcher, industrial firms that run A/B tests almost always immediately implement the interventions they are testing after an experiment. This means industrial analysts have a much more *consistent* and *shorter* feedback loop between statistical analysis and real-world implementation. This may give analysts the ability to quickly learn how p -hacking can lead to spurious results through first-hand experience, and allow them to adjust their behavior accordingly. We do not claim to have proven that these factors definitively explain our findings, but—particularly in contrast to prior literature on this subject—we believe our results reinforce the importance of (i) economic incentives and (ii) short feedback loops between experimentation and application for improving statistical rigor across institutional settings.

5.3 Conclusion. We conclude with several managerial implications of our project. 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. Though we cited many articles in this paper warning about the perils of p -hacking in managerial decision-making, when we investigated this phenomenon rig-

ously in a large sample of e-commerce websites, we failed to find evidence of this behavior in our research context. Our finding may give managers a measure of confidence in using and interpreting classical A/B test results to inform consequential decisions. In particular, though sophisticated experts recommend the use of more advanced statistical techniques from the sequential testing literature (such as the sequential probability ratio test) to prevent the negative effects of optional stopping or data peeking, our results suggest this extra effort may not be necessary (Larsen et al., 2024). If we accept our earlier conclusions about the importance of economic incentives and organizational learning processes, our findings suggest that these factors may be sufficient to encourage statistical rigor in data-driven decision making. Rather than focusing solely on technical solutions, managers might consider prioritizing the development of an organizational culture that aligns incentives with sound methodological practices. These conclusions would be consistent with and underscore existing advice given in the popular industry handbook *Trustworthy Online Controlled Experiments* (in Chapter 4 on “Experimentation Platform and Culture”), in which the authors suggest that managers should “[improve] agility with short release cycles to create a healthy, quick feedback loop for experimentation”, and that “[t]hey must provide the right incentives, processes, and empowerment for the organization to make data-driven decisions”. (Kohavi et al., 2020).

Perhaps the clearest implication of our research is that *before* making dramatic changes to experimentation practices—whether it be adopting a new statistical framework or retraining analysts 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 such proprietary datasets. In total, we believe this project serves as an important contribution to the literature on the industrial practice of statistical science.

References.

- 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.
- Andrews, I. and Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–94.
- Anscombe, F. J. (1954). Fixed-sample-size analysis of sequential observations. *Biometrics*, 10(1):89–100.
- 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.
- 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>.
- Berman, R. and Van den Bulte, C. (2022). False discovery in A/B testing. *Management Science*, 68(9):6762–6782.
- Borden, P. (2014). How Optimizely (Almost) Got Me Fired. <https://web.archive.org/web/20180608142925/https://blog.sumall.com/got-me-fired.html>.
- Brodeur, A., Carrell, S., Figlio, D., and Lusher, L. (2023a). Unpacking p-hacking and publication bias. *American Economic Review*, 113(11):2974–3002.
- Brodeur, A., Cook, N., and Heyes, A. (2020). Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review*, 110(11):3634–3660.
- Brodeur, A., Cook, N., and Heyes, A. (2023b). We Need to Talk About Mechanical Turk: What 22,989 Hypothesis Tests Tell Us About Publication Bias and P-Hacking in Online Experiments. *IZA Discussion Paper No. 15478*. Available at SSRN: <https://ssrn.com/abstract=4188289> or <http://dx.doi.org/10.2139/ssrn.4188289>.
- 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>.
- Cattaneo, M. D., Jansson, M., and Ma, X. (2018). Manipulation testing based on density discontinuity. *The Stata Journal*, 18(1):234–261.
- Chickering, D. M. and Heckerman, D. (2000). A decision theoretic approach to targeted advertising. *Uncertainty in Artificial Intelligence Proceedings*.
- Christian, B. (2012a). The A/B Test: Inside the Technology That’s Changing the Rules of Business. *WIRED*. <https://www.wired.com/2012/04/ff-abtesting/>.
- Christian, B. (2012b). Test Everything: Notes on the A/B Revolution. *WIRED*. <https://www.wired.com/2012/05/test-everything/>.
- Cox, D. R. and Hinkley, D. V. (1979). *Theoretical statistics*. CRC Press.
- 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.
- Draper, P. (2016). The Fatal Flaw of A/B Tests: Peeking. <https://www.lucidchart.com/blog/the-fatal-flaw-of-ab-tests-peeking>.
- 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.
- Feng, E. (2017). Building an Intelligent Experimentation Platform with Uber Engineering. <https://eng.uber.com/experimentation-platform/>.
- Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd; Edinburgh.
- Fiske, D. W. and Jones, L. V. (1954). Sequential analysis in psychological research. *Psychological Bulletin*, 51(3):264.
- Flory, J. (2021). The top 3 mistakes that make your A/B test results invalid. <https://www.widerfunnel.com/blog/3-mistakes-invalidate-ab-test-results/>.
- Gamber, T. (2019). Making Sense of A/B Testing Statistics. <https://www.confidenceinterval.com/blog/making-sense-of-ab-testing-statistics/>.
- 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.
- Gerber, A., Malhotra, N., et al. (2008). Do statistical reporting standards affect what is published? Publication

- bias in two leading political science journals. *Quarterly Journal of Political Science*, 3(3):313–326.
- Gerber, A. S. and Malhotra, N. (2008). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research*, 37(1):3–30.
- Ghosh, S., Thomke, S., and Pourkhalkhali, H. (2020). The effects of hierarchy on learning and performance in business experimentation. In *Academy of Management Proceedings*, volume 2020, page 20500. Academy of Management Briarcliff Manor, NY 10510.
- Hall, T. A. and Hasan, S. (2020). The Politics of Experimentation. *Available at SSRN 3571296*.
- Hamermesh, D. S. (2013). Six decades of top economics publishing: Who and how? *Journal of Economic Literature*, 51(1):162–172.
- 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.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8):e124.
- Johari, R., Koomen, P., Pekelis, L., and Walsh, D. (2017). Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1517–1525.
- 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. (2018). P-hacking in A/B Testing Sensationalized. <https://www.linkedin.com/pulse/p-hacking-ab-testing-sensationalized-ronny-kohavi/>.
- 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.
- Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181.
- Kohavi, R., Tang, D., and Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press.
- Kohavi, R. and Thomke, S. (2017). The surprising power of online experiments. *Harvard business review*, 95(5):74–82.
- 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.
- Larsen, N., Stallrich, J., Sengupta, S., Deng, A., Kohavi, R., and Stevens, N. T. (2024). Statistical challenges in online controlled experiments: A review of a/b testing methodology. *The American Statistician*, 78(2):135–149.
- Leahey, E. (2005). Alphas and asterisks: The development of statistical significance testing standards in sociology. *Social Forces*, 84(1):1–24.
- Lewis, R. A. and Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising. *The Quarterly Journal of Economics*, 130(4):1941–1973.
- 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.
- McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, 142(2):698–714.
- McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley Sons.
- McShane, B. B. and Gal, D. (2016). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science*, 62(6):1707–1718.
- Miller, E. (2010). How Not To Run an A/B Test. <http://www.evanmiller.org/how-not-to-run-an-ab-test.html>.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.
- Overgoor, J. (2014). Experiments at Airbnb. <https://medium.com/airbnb-engineering/>

experiments-at-airbnb-e2db3abf39e7.

- 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). <https://blog.optimizely.com/2015/01/20/statistics-for-the-internet-age-the-story-behind-optimizelys-new-stats-engine/>.
- 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.
- 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.
- Shaw, S. D. and Nave, G. (2023). Don’t hate the player, hate the game: Realignment incentive structures to promote robust science and better scientific practices in marketing. *Journal of Business Research*, 167:114129.
- 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.
- Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2013). Life after p-hacking. In *Meeting of the society for personality and social psychology*, New Orleans, LA, pages 17–19.
- 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.
- Spiess, J. (2018). Optimal estimation when researcher and social preferences are misaligned.
- Stanley, T. D. and Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1):60–78.
- Szucs, D. (2016). A tutorial on hunting statistical significance by chasing N. *Frontiers in psychology*, 7:1444.
- Tambe, P. and Hitt, L. M. (2014). Job hopping, information technology spillovers, and productivity growth. *Management science*, 60(2):338–355.
- Tang, D., Agarwal, A., O’Brien, D., and Meyer, M. (2010). Overlapping Experiment Infrastructure: More, Better, Faster Experimentation. In *Proceedings 16th Conference on Knowledge Discovery and Data Mining*, pages 17–26, Washington, DC.
- Thomke, S. H. (2020). *Experimentation works: The surprising power of business experiments*. Harvard Business Press.
- 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/>.
- Vogel, D. and Homberg, F. (2021). P-Hacking, P-Curves, and the PSM–Performance Relationship: Is There Evidential Value? *Public Administration Review*, 81(2):191–204.
- Wald, A. (1945). Sequential tests of statistical hypotheses. *The annals of mathematical statistics*, 16(2):117–186.
- Walker, T. (2015). Warning: Most Conversion Optimization Tips Are BS (Here’s Why!). <https://www.shopify.com/enterprise/44310083-warning-most-conversion-optimization-tips-are-bs-heres-why>.
- Warwick, M. (2003). *Testing, Testing 1, 2, 3: Raise More Money with Direct Mail Tests*. John Wiley & Sons.

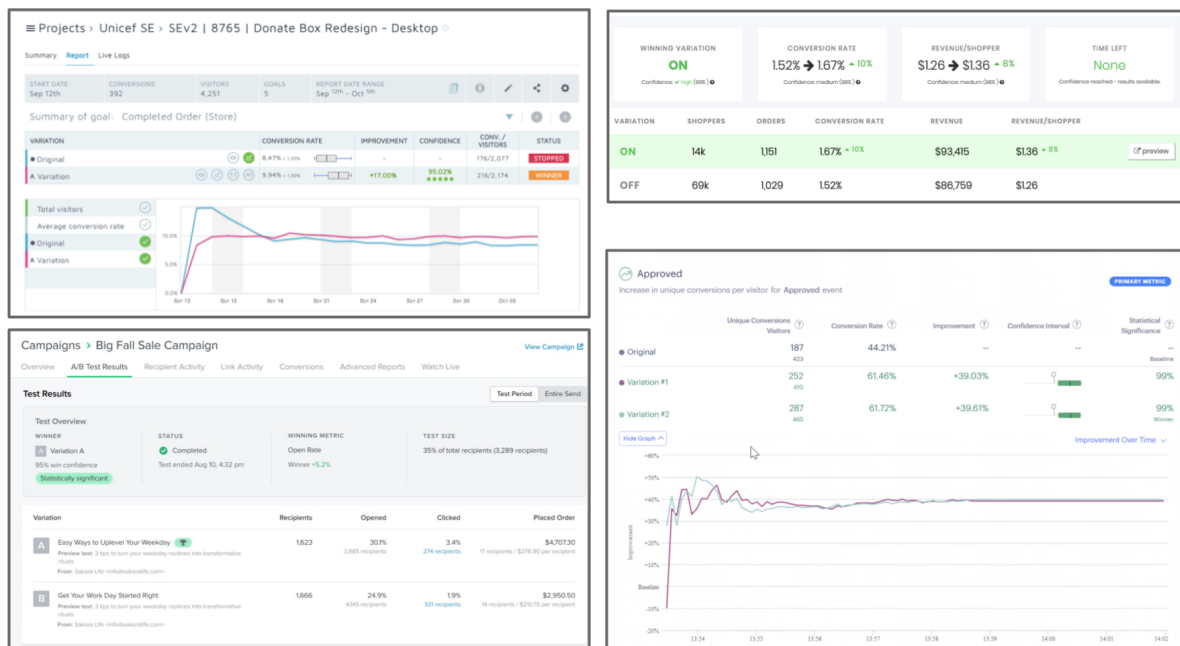
Supplementary material to accompany the manuscript

“An investigation of p -hacking in e-commerce A/B testing”

APPENDICES.

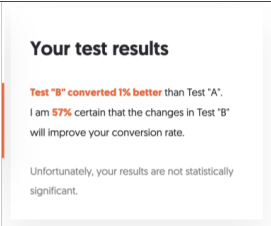
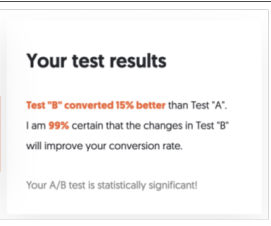
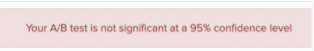

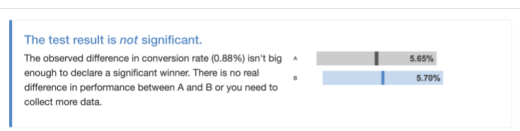
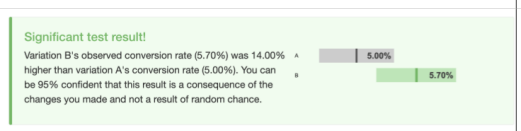
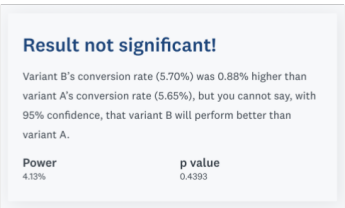
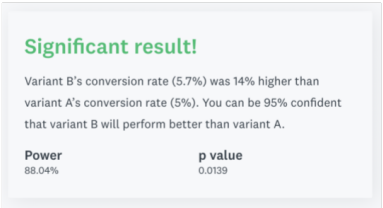
A Common design paradigms in A/B testing software.

Figure A.1: Examples of industry dashboards used to report the results of A/B tests



Notes: Screenshots of dashboards taken from real A/B testing software interfaces. Clockwise from top-left, the platforms are: Convert.com, Klayvio, Fera.ai, and Optimizely.

Figure A.2: How online A/B testing tools describe results that reach 95% “confidence”

| | | |
|-----|--|--|
| (a) |  <p>Your test results</p> <p>Test "B" converted 1% better than Test "A". I am 57% certain that the changes in Test "B" will improve your conversion rate.</p> <p>Unfortunately, your results are not statistically significant.</p> |  <p>Your test results</p> <p>Test "B" converted 15% better than Test "A". I am 99% certain that the changes in Test "B" will improve your conversion rate.</p> <p>Your A/B test is statistically significant!</p> |
| (b) |  <p>Your A/B test is not significant at a 95% confidence level</p> |  <p>The winning variation is Variation B at a 95% confidence level.</p> |
| (c) |  <p>The test result is <i>not</i> significant.</p> <p>The observed difference in conversion rate (0.88%) isn't big enough to declare a significant winner. There is no real difference in performance between A and B or you need to collect more data.</p> |  <p>Significant test result!</p> <p>Variation B's observed conversion rate (5.70%) was 14.00% higher than variation A's conversion rate (5.00%). You can be 95% confident that this result is a consequence of the changes you made and not a result of random chance.</p> |
| (d) |  <p>Result not significant!</p> <p>Variant B's conversion rate (5.70%) was 0.88% higher than variant A's conversion rate (5.65%), but you cannot say, with 95% confidence, that variant B will perform better than variant A.</p> <p>Power 4.13%</p> <p>p value 0.4393</p> |  <p>Significant result!</p> <p>Variant B's conversion rate (5.7%) was 14% higher than variant A's conversion rate (5%). You can be 95% confident that variant B will perform better than variant A.</p> <p>Power 88.04%</p> <p>p value 0.0139</p> |

Notes: The images above are screenshots from online “significance calculators”, which are designed to help analysts understand results from an A/B test. In the first column, we used each tool to calculate the results of an imaginary experiment with 10,000 observations in each arm with 565 conversions observed in arm “A” and 570 conversions in arm “B”; in the second column, we decreased the number of conversions in arm “A” to 500 and had the tool recalculate the test’s results. Sources are as follows: (a) <https://neilpatel.com/ab-testing-calculator/>, (b) <https://www.convertize.com/ab-test-significance/>, (c) <https://abtestguide.com/calc/>, (d) <https://www.surveymonkey.com/mp/ab-testing-significance-calculator/>

B Asymmetric caliper test for discontinuity detection.

In this article’s main text, we laid out a stripped-down outline of the asymmetric caliper test we developed for this project. In this appendix, we provide more detailed information on the formalities of the statistical test and also provide evidence for why this test is preferable to other methods in our research context.

B.1 Test procedure, derivation of formulae, and statistical formalities. Assume we have a vector of N observed p -values denoted by x . We are interested in determining whether there is a disproportionate number of p -values just below a chosen threshold of interest τ (in our primary analysis, this is the conventional significance threshold of 0.05). Assume we have a given bandwidth value $h > 0$, which specifies the local region around τ in which we will be focusing our analysis.

We first formally define the empirically observed difference in the raw counts of p -values on either side of the specified threshold as follows:

$$S^* = \sum_{i=1}^N \mathbb{I}\{x_i \in [\tau - h, \tau)\} - \sum_{i=1}^N \mathbb{I}\{x_i \in [\tau, \tau + h)\}$$

Our test procedure is based on comparing this empirically observed statistic to the distribution of this value that is implied by assuming, as our null hypothesis, that our data x are adequately parameterized by the fully continuous mixture model f described in Section 4.3.

B.1.1 Null distribution of test statistic. Under the assumption of this null hypothesis, we can define a random variable S (conditional on N , τ , and h) as the difference in the number of observations above and below the threshold of interest. From a frequentist perspective, for a given density f and specified values of N , τ , and h , this quantity S is random as it can vary across different samples of data drawn from f . To describe this quantity formally, we begin by defining π_l as the fraction of the distribution in the window of width h on the *left* side of the critical threshold, and π_r as the fraction on the *right* side:

$$\pi_l = \int_{\tau-h}^{\tau} f(x)dx, \quad \pi_r = \int_{\tau}^{\tau+h} f(x)dx \quad (2)$$

(A visualization of these parameters is provided in Figure 8(a).) 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 :

$$n_l \sim \text{Binomial}(N, \pi_l), \quad n_r \sim \text{Binomial}(N, \pi_r)$$

We then formally define our test statistic as $S := n_l - n_r$, whose mass function can be derived by summing up all ways in which these two binomials can differ by some value $k \in \mathbb{N}$. This formula,

given below, defines the distribution of our test statistic under the null hypothesis:

$$\begin{aligned}
\phi(k) &:= P[S=k] \\
&= \begin{cases} \sum_{i=0}^N P[n_l=i+k] P[n_r=i], & \text{if } k \geq 0 \\ \sum_{i=0}^N P[n_l=i] P[n_r=i+k], & \text{otherwise} \end{cases} \\
&= \begin{cases} \sum_{i=0}^N \binom{N}{i+k} (\pi_l)^{i+k} (1-\pi_l)^{N-(i+k)} \binom{N}{i} (\pi_r)^i (1-\pi_r)^{N-i}, & \text{if } k \geq 0 \\ \sum_{i=0}^N \binom{N}{i} (\pi_l)^i (1-\pi_l)^{N-i} \binom{N}{i+k} (\pi_r)^{i+k} (1-\pi_r)^{N-(i+k)}, & \text{otherwise} \end{cases}
\end{aligned} \tag{3}$$

B.1.2 Maximum likelihood estimates of test statistic parameters. The distribution of the test statistic given in Eq. 3 has only been defined for a given density function f , parameterized by some known vector θ . In practice, θ and f must be estimated from the observed data x . We accomplish this in our project using maximum likelihood estimation. Recall that we have based our model for f on the well-established beta-uniform mixture model, which is commonly used in other research settings that deal with distributions of p -values. This model consists of K mixture components indexed by $k=0,1,\dots,K$, where the constituent parameters of $\theta=(\pi,\alpha,\beta)$ are described in Section 4.3.1.

To proceed, consider a single draw x_i from this model and integrate over our uncertainty in the mixture component to which i belongs; this gives us an expression for the marginal density of f :

$$f(x_i|\theta) = \int_k f(x_i|\theta, k_i=k) dF(k) = \pi_0 + \sum_{k=1}^K \pi_k f_k(x_i) \tag{4}$$

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 :

$$\mathcal{L}(\theta|x) = \prod_{i=1}^N f(x_i|\theta) = \prod_{i=1}^N \left(\pi_0 + \sum_{k=1}^K \pi_k f_k(x_i) \right) \tag{5}$$

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

$$\begin{aligned}
\ell(\theta|x) &= \log \mathcal{L}(\theta|x) \\
&= \log \left(\prod_{i=1}^N \left[\pi_0 + \sum_{k=1}^K \pi_k f_k(x_i) \right] \right) \\
&= \sum_{i=1}^N \log \left(\pi_0 + \sum_{k=1}^K \pi_k f_k(x_i) \right)
\end{aligned} \tag{6}$$

For a given dataset x , an empirical estimate of our model's primary parameters is then given by

$$\hat{\theta} = \operatorname{argmax}_{\theta} \ell(\theta|x).$$

We can now define the maximum likelihood estimate for the null distribution of our test statistic $\hat{\phi}$ (defined in Eq. 3) by first deriving estimates of $\hat{\pi}_l$ and $\hat{\pi}_r$ (defined in Eq. 2). Conditional on estimated model parameters, this can be done in closed form:

$$\begin{aligned}
\hat{\pi}_r &= \int_{\tau}^{\tau+h} \hat{f}(x) dx \\
&= \int_{\tau}^{\tau+h} \left[\hat{\pi}_0 + \sum_{k=1}^K \hat{\pi}_k \hat{f}_k(x) \right] dx \\
&= \pi_0 \int_{\tau}^{\tau+h} dx + \sum_{k=1}^K \pi_k \int_{\tau}^{\tau+h} \hat{f}_k(x) dx \\
&= \hat{\pi}_0 h + \sum_{k=1}^K \hat{\pi}_k \left[I(\tau+h; \hat{\alpha}_k, \hat{\beta}_k) - I(\tau; \hat{\alpha}_k, \hat{\beta}_k) \right]
\end{aligned} \tag{7}$$

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

B.1.3 Estimation of model parameters by means of expectation-maximization. Our mixture model, given the presence of latent unobservable component assignment variables, lends itself well to optimization via expectation-maximization (McLachlan and Peel, 2000). We use a version of the E-M algorithm in an iterative fashion to estimate the parameters of our model; pseudo-code for the process is given below:

Algorithm 1: Expectation-Maximization for Beta-Uniform Finite Mixture Model

Input : Observed data $\mathbf{x} = (x_1, \dots, x_n)$, number of beta components K , initial parameters $\pi^{(0)}, \alpha^{(0)}, \beta^{(0)}$

Output: Estimated parameters π, α, β

- 1 Define $f_k(x; \alpha_k, \beta_k)$ as the PDF of the k -th Beta component;
- 2 $\theta^{(0)} \leftarrow (\pi^{(0)}, \alpha^{(0)}, \beta^{(0)})$;
- 3 **while** not converged **do**
- 4 **E-step:** Compute the expected latent variables given the observed data and current parameters;
- 5 **for** $i \leftarrow 1$ **to** n **do**
- 6 **for** $k \leftarrow 1$ **to** K **do**
- 7 $\gamma_{ik}^{(t+1)} \leftarrow \frac{\pi_k^{(t)} \cdot f_k(x_i; \alpha_k^{(t)}, \beta_k^{(t)})}{\pi_0^{(t)} + \sum_{k=1}^K \pi_k^{(t)} \cdot f_k(x_i; \alpha_k^{(t)}, \beta_k^{(t)})}$;
- 8 **end**
- 9 **end**
- 10 $\pi_0^{(t+1)} \leftarrow 1 - \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \gamma_{ik}^{(t+1)}$
- 11 **M-step:** Maximize the expected complete-data log-likelihood with respect to α and β , holding π fixed;
- 12 Derive π from γ by averaging across sample: $\pi_k^{(t+1)} \leftarrow \frac{1}{n} \sum_{i=1}^n \gamma_{ik}^{(t+1)}$;
- 13 $\alpha^{(t+1)}, \beta^{(t+1)} \leftarrow \operatorname{argmax}_{\alpha, \beta} \ell(\alpha, \beta | \pi^{(t+1)}, \mathbf{x})$;
- 14 Update the parameter estimates: $\theta^{(t+1)} \leftarrow (\pi^{(t+1)}, \alpha^{(t+1)}, \beta^{(t+1)})$;
- 15 **end**
- 16 **return** π, α, β ;

B.1.4 Definition of p -value for test statistic. Having derived the null distribution of our test statistic and described a procedure for estimating this from our data, we can now derive a p -value for our analysis. This value will represent the probability that we would observe a test statistic (i.e., the difference between the number of test results just above and below the critical threshold) as large or larger than the observed statistic, assuming the sampling distribution of the test statistic follows the formula derived in Eq. 3, with parameters estimated via the MLE procedure described in Section B.1.2.

Because p -hacking behavior is only consistent with observing an excess mass of p -values *below* the critical threshold, we will make use of a *one-sided* p -value, defined as follows:

$$p\text{-value (one-sided)} = P[S^* \leq \hat{S}] = \sum_{k=S^*}^{\infty} P[\hat{S}=k] \quad (8)$$

For completeness, a two-sided p -value can be defined analogously. Even though the distribution of our null statistic is technically not guaranteed to be symmetric about its mode, we can still closely approximate the exact two-sided p -value by simply doubling the appropriate one-sided p -value.¹³ If we define the two-sided p -value as the probability that the null test statistic would be as or more “extreme” than the observed statistic, then the direction of “extremity” used to define our base one-sided p -value depends on whether the observed statistic is above or below the center of the null distribution. Specifically, we use the following formula for this approximation:

$$p\text{-value (two-sided)} = \begin{cases} \text{if } S^* > \mathbf{E}[\hat{S}]: & 2P[S^* \leq \hat{S}] = \sum_{k=S^*}^{\infty} P[\hat{S}=k] \\ \text{if } S^* \leq \mathbf{E}[\hat{S}]: & 2P[\hat{S} \leq S^*] = \sum_{k=-\infty}^{S^*} P[\hat{S}=k] \end{cases} \quad (9)$$

B.1.5 Definition of confidence intervals. Following the definition of Cox and Hinkley (1979), for a given significance level α , we can derive a $100 \times (1 - \alpha)\%$ confidence interval by computing all values $s \in \mathbb{N}$ for which we cannot reject the null hypothesis that the difference between our observed and null test statistics is exactly s :

$$C(S^*) = \left\{ s \in \mathbb{N} \mid H_0 : S^* - \hat{S} = s \text{ is not rejected at significance level } \alpha \right\} \quad (10)$$

This single definition works for both one- and two-sided confidence intervals, where one merely needs to adapt either the one- or two-sided p -values defined in Eqns. 8 and 9 to determine whether or not H_0 in Eq. 10 is rejected for various values of s .

¹³Recall our test statistic is based on the difference of two binomially-distributed random variables; as N gets large, the binomial distribution is very well-approximated by the normal distribution, and the difference of two normal distributions is also normal—simplifying this computation dramatically. Further, because our outcomes of interest are in integer counts, only in rare circumstances will a confidence interval derived via this approximation differ from its exact value. For these reasons, we argue this simplification is justified in practice.

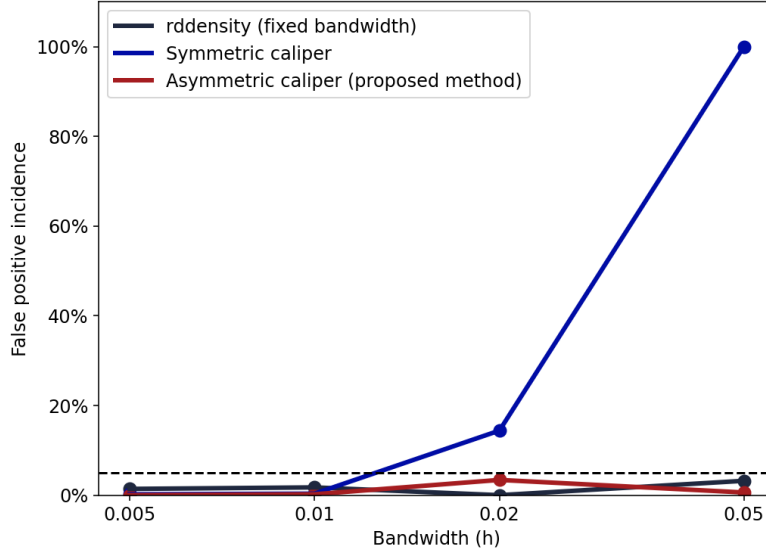
B.2 Comparison to existing methods. In this section, we compare our proposed technique described above to two other well-established methods for density discontinuity detection (see Section 4.2 in the main text for more context on different techniques):

- The “symmetric caliper” technique, as first applied to study the effect of significance thresholds in Gerber and Malhotra (2008). This method consists of a simple binomial test of equal proportions. For a given threshold τ and bandwidth h , this tests whether the number of observations that fall below the threshold, in the interval $[\tau-h, \tau)$, exceeds 50% of the total number of observations that fall within the entire window $[\tau-h, \tau+h)$.
- The local polynomial technique of Cattaneo et al. (2018), **rddensity**, as implemented in their R package.

B.2.1 Analysis of false positive rates. In our first analysis, we attempted to assess how these methods perform in terms of their false positive rate. Importantly, because the bandwidth h is a researcher-specified parameter with no objective value, it is important for a test of discontinuity to be robust to different values of this parameter. To assess this, we generated 1,000 samples of p -values from a beta-uniform mixture model, with parameters set to their values fit to our dataset by MLE. (This guarantees that the underlying distribution of p -values is indeed continuous; however, because there appears to be no discontinuity in our raw data, the results we report below are essentially identical if we use simple bootstrap sampling from empirically observed p -values.) Then for each bandwidth value $h \in \{0.005, 0.01, 0.02, 0.05\}$, we use all three methods described above to perform a statistical test for discontinuity at the $\tau = 0.05$ threshold, with the bandwidth window for each test set to h . In each case, we rejected the null hypothesis if the p -value from a test was below $\alpha = 0.05$. In Figure A.3, we plot the proportion of times each method rejects the null hypothesis—all of which are, by construction, false positives. We have plotted a dashed line at $\alpha = 0.05$ to facilitate comparison with the tests’ nominal false positive control rate.

As can be seen, both our proposed technique and the non-parametric **rddensity** technique adequately control the false positive rate below the 5% level for all values of the bandwidth parameter. However, we can see that starting at $h = 0.02$, the false positive rate of the symmetric caliper technique jumps to 14.4%. Because of the extreme skew of p -value distributions, when the bandwidth spans all the way to zero (i.e., $[\tau-h, \tau) = [0, 0.05)$), the symmetric caliper rejects 100% of the tests for discontinuity since it assumes that the proportion of observations on either side of the $\tau = 0.05$ threshold is equal. Because our method is designed specifically to account for this shape in the distribution of p -values, the *asymmetric* caliper is able to control its positive rate below the desired level.

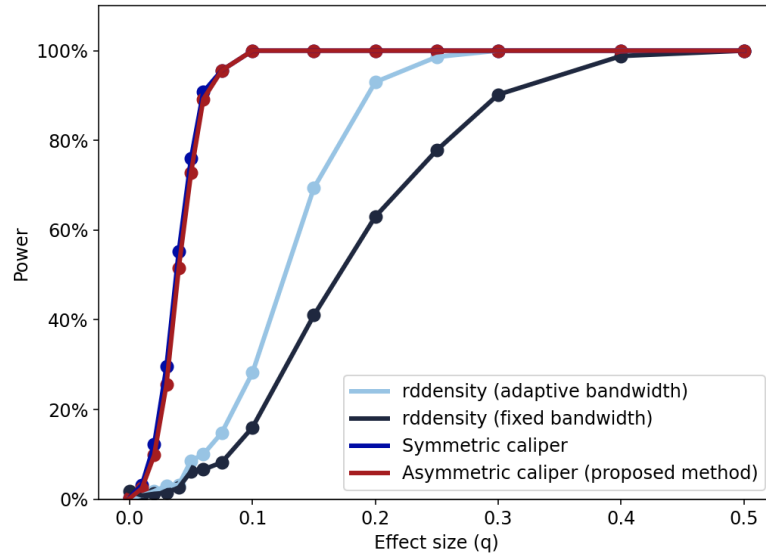
Figure A.3: False rejection rate for discontinuity at $\tau=0.05$, with $\alpha=0.05$



B.2.2 Power analysis. All statistical tests make a trade-off between false-positive control and power—i.e., the ability to detect an effect when one is truly present. One might be concerned that while our technique appears to do well on the former objective, it may falter in performance on the latter. To assess this, we performed counterfactual simulations in which we simulated daily time-series of p -values that would be observed throughout the lifetime of an experiment and then p -hacked a proportion q of experiments in each simulation (using the technique described in Appendix C). This was done 1,000 times for each of 15 different values of q in the range $[0,0.5]$. Then we performed four different tests for discontinuity at this threshold: the same three tests analyzed above with the bandwidth set to $h=0.01$, as well as the adaptive bandwidth version of **rddensity**, which non-parametrically estimates an optimal bandwidth.

The results of this analysis are shown in Figure A.4, in which we plot the simulated power curve for all four tests. As can be seen, the symmetric and asymmetric versions of the caliper test have very similar power curves. The symmetric caliper has slightly higher power for small effect sizes but, as discussed above, this comes at the cost of a potentially inflated false positive rate (that is sensitive to the underlying nuisance parameter, h). Our method appears to achieve both significantly higher power than the **rddensity** methods while simultaneously controlling for false positives across a range of bandwidth values. Given this analysis, we argue there is strong evidence that the asymmetric caliper is able to strike a better balance between sensitivity and specificity than the other methods considered here.

Figure A.4: Power analysis



Note: Power is computed as the number of times each test rejects the null hypothesis out of 1,000 simulations for each q , wherein each simulation, we draw a bootstrap sample of size N tests from our dataset and p -hack a fraction q of them (see Appendix C for details). For all tests except the adaptive bandwidth version of **rddensity**, the bandwidth is set to $h = 0.01$.

C Counterfactual Monte Carlo simulation for power analysis.

Algorithm 2: Counterfactual p -hacking and power calculation 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) of discontinuity test at chosen α level, averaged by bootstrap

Fixed variables:

E , dataset of experiments, each experiment consists of the terminal outcomes

observed for the experiment in our dataset (e.g., number of conversions and sessions for each arm)

$B \gg 0$, bootstrap sample size ;

Notation:

T_i , empirically observed duration of experiment i in days;

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 $\{1, \dots, Z\}$ for $Z \in \mathbb{N}_+$;

$\mathcal{S}(\cdot)$, statistical test for discontinuity, as described in Appendix B (returns p -value);

\mathcal{P} , set of p -values for statistical test, observed across bootstrap samples

```

1  function CounterfactuallyPHackExperiment (Experiment  $i$ )
2      DailyData $_i \leftarrow$  Simulate daily data for each arm in experiment
        by distributing each arm's terminal observations multinomially across number of days in experiment  $T_i$ 
3       $\{p_{it} \mid i \in [T_i]\} \leftarrow$  Calculate daily  $p$ -values from DailyData $_i$ 
4      Set  $p_i^*$  to NULL
5      for day  $t \in [T_i]$  do
6          if  $p_{it} < 0.05$  then
7               $p_i^* \leftarrow p_{it}$                                 set to  $p$ -hacked value
8              break loop; go to 19
9      if  $p_i^*$  is NULL then
10         for day  $t \in \{T_i + 1, \dots, 2T_i\}$  do
11             ExtraData $_{it} \leftarrow$  Simulate new data using Binomial distribution
                for binary outcomes (with parameter set to observed rate for each arm), and Poisson
                distribution for non-binary outcomes (with mean set to observed daily average for each arm)
12             DailyData $_i \leftarrow$  DailyData $_i \cup$  ExtraData $_{it}$ 
13              $p_{it} \leftarrow$  Calculate cumulative  $p$ -value using all data observed in DailyData $_i$ 
14             if  $p_{it} < 0.05$  then
15                  $p_i^* \leftarrow p_{it}$                                 set to  $p$ -hacked value
16                 break loop; go to 19
17             if  $t == \max\{30, 2T_i\}$  then
18                  $p_i^* \leftarrow p_{it}$                                 set terminal  $p$ -value to observed value, even if not significant
19         return  $p_i^*$ 
20 function PowerCalculation( $q, N$ )
21     for bootstrap  $b \in [B]$  do
22          $E_N \leftarrow$  random sample of size  $N$  from  $E$                                 with replacement
23          $P_b \leftarrow \{\}$ 
24          $H_b \leftarrow$  random sample of proportion  $q$  from  $E_N$ 
25         for experiment  $i \in E_N$  do
26             if  $i \in H_b$  then
27                  $p_i^* \leftarrow$  CounterfactuallyPHackExperiment( $i$ )
28             else
29                  $p_i^* \leftarrow p_{iT_i}$                                 use empirically observed terminal  $p$ -value
30                  $P_b \leftarrow \{p_i^*\} \cup P_b$                                 add simulated  $p$ -value to observation set
31          $\pi_b \leftarrow \mathcal{S}(P_b)$                                 apply statistical test for discontinuity to  $p$ -hacked  $p$ -values
32          $\mathcal{P} \leftarrow \{\pi_b\} \cup \mathcal{P}$                                 collect  $p$ -values of discontinuity test
33      $R \leftarrow \{1[p < \alpha] \text{ for } p \in \mathcal{P}\}$                                 indicator set of rejected tests
34     power  $\leftarrow \sum R / B$                                 average rejection proportion across bootstraps
35     return power

```

D Reasons for choosing a fixed rather than adaptive bandwidth parameter h .

When estimating density functions or dealing with other statistical smoothing problems, the selection of any particular bandwidth value h is ultimately arbitrary. The primary bandwidth used for our testing procedure in the body of the text is fixed at $h=0.01$; we also report the results of our analysis when $h=0.005$ and $h=0.02$ and find consistent results. To be sure, our results also remain consistent when h is set to any of the 20 logarithmically-spaced h -values between 0.001 and 0.10. However, a reasonable technique for choosing h that we omitted from our main text is to estimate an “optimal” value based on relevant statistical or information-theoretic criteria (Berman et al., 2018, Rudemo, 1982). Given the multiplicity of approaches, we find it incumbent to motivate our choice bandwidth parameters.

There is prior research that motivates the selection of $h=0.01$ in the context of studying anomalies near significance thresholds, which is why this is our default value Simonsohn et al. (2014). But we believe can motivate out of principle (at least in our context) why selecting an fixed value for h is preferable over using a data-adaptive approach. Indeed, we have attempted to use more data-adaptive methods for selecting the bandwidth value and find consistently that these methods suffer from erratic and unpredictable statistical behavior in small samples. These techniques are still appropriate for use with large N . However, at our sample size of around 2,300 observations, the data-adaptive approach is demonstrably 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:

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

The bandwidth selection problem can then be formally expressed as:

$$h^* = \operatorname{argmax}_h MISE(\hat{f}_h)$$

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:

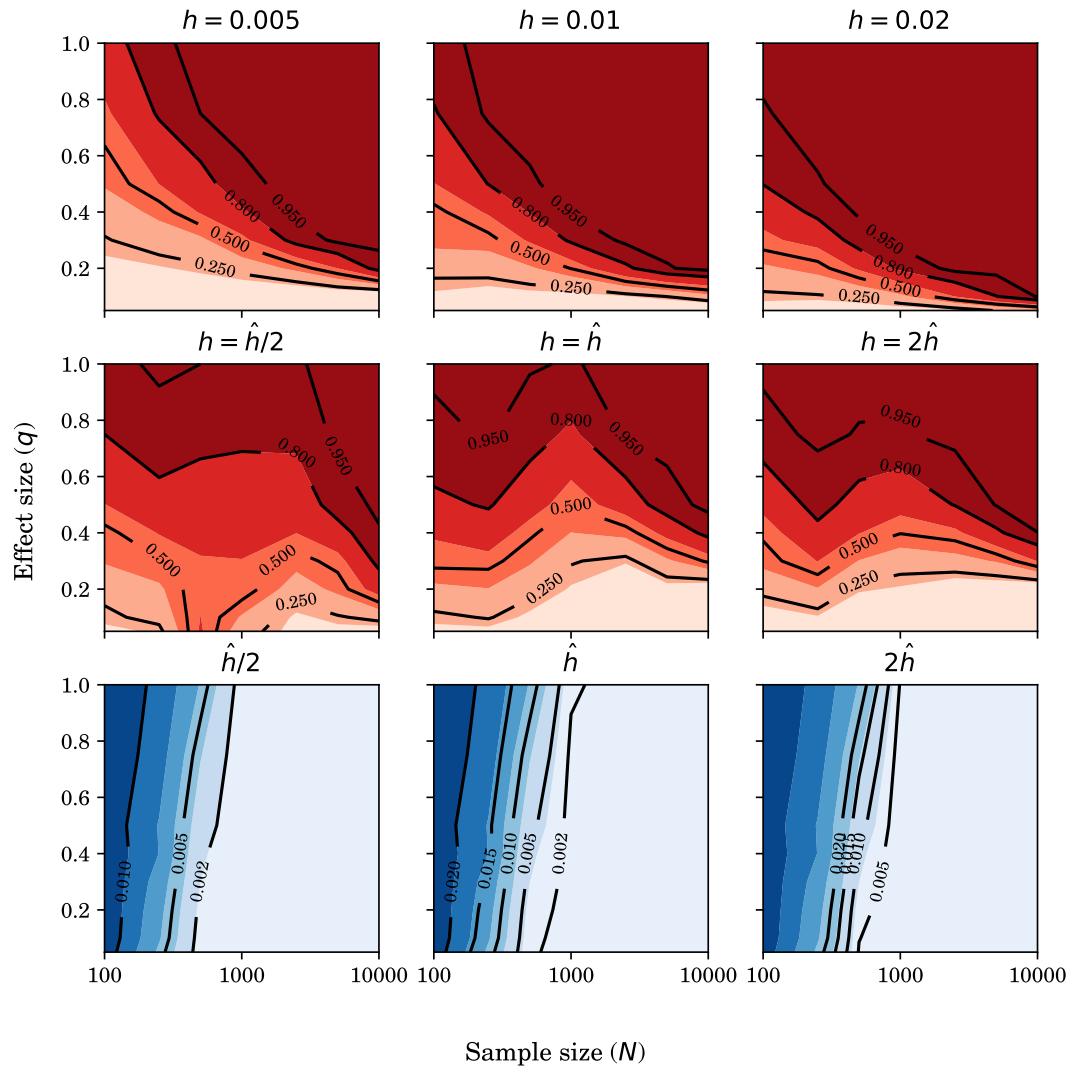
$$MISE(h) = \frac{2}{(n-1)h} - \frac{n+1}{n^2(n-1)h} \sum_k N_k^2$$

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). We are able to estimate the power of each approach empirically, using the counterfactual method described in Appendix C. Figure A.5 shows the results of our power analysis, where we have varied both the effect size (fraction of tests p -hacked, q) and the sample size of A/B tests used in our simulations (N). For a grid of values over these variables, we used either the fixed or adaptive values of selecting the optimal h . The first row contains the resulting power contours of 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 A.5 shows the average value of the estimated optimal 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 irregularities appear to mellow out for large samples sizes (near $N = 10,000$), the size of our empirical sample is less than 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 do not appear particularly sensitive to this choice.

Figure A.5: Effects of fixed vs. adaptive bandwidth values on statistical power



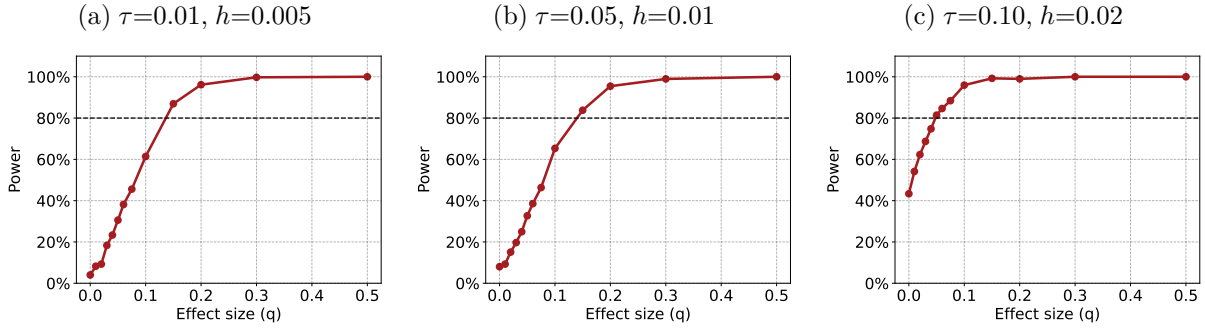
Note: Top row: estimated power of discontinuity detection procedure for **fixed bandwidths** across sample sizes (inner x -axes) and counterfactual effect sizes (inner y -axes); bandwidth values specified above plots. Middle row: estimated power of **adaptive bandwidth** estimators (with bandwidths set to half, full, and twice the optimal values, respectively). Bottom row: estimated optimal bandwidth, as determined by LOOCV criteria for histogram density estimation (center panel); left and right plots show half and double optimal bandwidth, respectively.

E Heterogeneous thresholds.

One possibility not considered in the main text is that different firms may p -hack their experiments at different thresholds. Assuming this were the case, we would expect to see weaker evidence of p -hacking at any given threshold which would make this behavior more difficult to detect. We can expand on the counterfactual simulation developed in Section 4.4.1 to determine if and how our proposed method for detecting p -hacking behavior would handle this scenario.

To this end, we altered our simulation procedure so that rather than assuming all firms were p -hacking at the same threshold, we allocate each firm one of the three thresholds in $\{0.01, 0.05, 0.10\}$. In each round of the simulation, a firm is randomly assigned one of these thresholds and then assume that if a firm intends to p -hack one of their experiments, they p -hack at their firm-specific threshold. We then proceed as before, assuming that a fraction q of the experiments in our data were intended to be p -hacked, and then simulate the effect of this behavior on the final distribution of p -values. At the end of each run of a simulation, we run our discontinuity test at each of the three thresholds. We then repeat this simulation 1000 times and keep track of how often our test is able to detect the simulated discontinuous behavior (assuming a rejection threshold of $\alpha=0.05$). We report the results of this simulation below and discuss the implications.

Figure A.6: Power for tests of discontinuity assuming firms p -hack at different thresholds



Note: Discontinuity detection applied at τ with bandwidth h .

For an apples-to-apples comparison with the power results reported in the main text, we first highlight the results in Figure A.6(b). Our results here suggest we would have 80% power to detect a discontinuity at $\tau=0.05$ assuming an effect size of approximately $q \approx 0.12$ (i.e., assuming 12% of the experiments were intended to be p -hacked). We also see a similar power curve at the $\tau=0.01$ threshold. For the $\tau=0.10$ threshold, we find that we should be able to detect relatively small effects, on the order of $q=0.05$, with 80% power. We conclude from this that the effects of heterogeneous thresholds would be to reduce our test's power at the lower thresholds, but assuming that some portion of firms were p -hacking at the $\tau=0.10$ threshold, our simulations suggest we should still be relatively well-powered to observe small effects ($q \approx 0.05$). Overall, the effects of

heterogeneity may reduce our test's power somewhat, but we believe these results are by and large consistent with the conclusions of our main analysis: Assuming there was a meaningful amount of experiments that were intended to be p -hacked, this behavior would result in discontinuities in the aggregate distribution of p -values, and these discontinuities would be observable and detectable with the techniques used in this project. To be sure, the aforementioned procedure is likely too formulaic to reflect true human behavior. Nonetheless, this exercise is useful for helping understand the robustness of our methods to more complex behavior than what was assumed in our main text and provides a useful extension of our primary analysis.