p-Hacking and False Discovery in A/B Testing

Ron Berman*

Leonid Pekelis†

Aisling Scott‡

Christophe Van den Bulte§

December 11, 2018

Abstract

We investigate to what extent online A/B experimenters "p-hack" by stopping their experiments based on the p-value of the treatment effect, and how such behavior impacts the value of the experimental results. Our data contains 2,101 commercial experiments in which experimenters can track the magnitude and significance level of the effect every day of the experiment. We use a regression discontinuity design to detect the causal effect of reaching a particular p-value on stopping behavior. Experimenters indeed p-hack, at times. Specifically, about 73% of experimenters stop the experiment just when a positive effect reaches 90% confidence. Also, approximately 75% of the effects are truly null. Improper optional stopping increases the false discovery rate (FDR) from 33% to 40% among experiments p-hacked at 90% confidence. Assuming that false discoveries cause experimenters to stop exploring for more effective treatments, we estimate the expected cost of a false discovery to be a loss of 1.95% in lift, which corresponds to the 76^{th} percentile of observed lifts.

We thank Optimizely for making the data available for this project. We also thank Eric Bradlow, Elea McDonnell Feit, Matthew Gershoff, Raghuram Iyengar, Pete Koomen, Lukas Vermeer and audiences at the University of Mannheim, the 2018 INFORMS Marketing Science Conference and an NYC Marketing Modelers Meeting for comments.

*Assistant Professor of Marketing, The Wharton School. ronber@wharton.upenn.edu

†Statistician, OpenDoor. leo@opendoor.com

‡Affiliation withheld. aisling_scott@haas.berkeley.edu

§Gayfryd Steinberg Professor and Professor of Marketing, The Wharton School. vdbulte@wharton.upenn.edu

1 Introduction

Marketers increasingly use online experiments to inform their decisions. This shift is facilitated by various testing platforms like Conductrics, Google Optimize, Mixpanel, Monetate and Optimizely. These A/B testing platforms, as they are often called, make it easy to randomly allocate consumers to treatment conditions and to measure their responses.

To help marketers avoid rolling out non-effective marketing treatments, platforms typically report standard null hypothesis significance tests. However, industry observers have pointed to the need to use these tests properly. Platforms calculate and display both effect sizes and significance levels on an ongoing basis. Users do not have to pre-commit to ending the experiment at a particular time or sample size, but can stop the experiment at any time. This ability to both (1) peek at the test results obtained while the experiment is still running (repeated testing) and (2) terminate the experiment at any time (optional stopping), may lead users to end the experiment based on the significance level achieved.

Such stopping behavior would be problematic, since repeated traditional null hypothesis significance testing over the course of the experiment inflates the false-positive rate. This has long been known (e.g., Wald 1947, Fiske and Jones 1954, Armitage et al. 1969, Bartroff et al. 2012), but was until recently ignored by most providers and users of A/B testing platforms (Miller 2010, Goodson 2014, Johari et al. 2017). In short, the combination of optional stopping and uncorrected repeated testing raises the specter that users might stop their experiment and declare their findings once the experiment reaches a desired statistical significance level. Stopping A/B tests in such a way would result in rejecting the null hypotheses more often than the nominal rate, inflate the number of false positives, and be a form of p-hacking.

p-Hacking is typically defined as "try[ing] out several statistical analyses and/or data eligibility specifications and then selectively report[ing] those that produce significant results" (Head et al. 2015, p. 1) or "the misuse of data analysis to find patterns in data that can be presented as statistically significant when in fact there is no real underlying effect".[1] Conducting statistical tests midway through experiments and using the test results to decide whether to continue collecting data without acknowledging, or correcting for, having tested the same hypothesis repeatedly is a specific form of p-hacking. Put more succinctly, improper optional stopping is a form of p-hacking, as

[1] `https://en.wikipedia.org/wiki/Data_dredging`, Accessed Nov 19, 2018.

many have stated explicitly (Rouder 2014, Head et al. 2015, Bruns and Ioannidis 2016, Szucs 2016, Brodeur et al. 2018). Importantly, p-hacking in general and through improper optional stopping specifically need not result from malicious intent to deceive, and more likely stems from an honest misunderstanding of statistical testing (Simmons et al. 2011, John et al. 2012, Yu et al. 2014). We use the term p-hacking without attributing or implying specific intentions to either platforms or users.

In response to these concerns about repeatedly checking for significance, A/B testing platforms have made substantial efforts to educate marketers about the importance of proper statistical inference. Some have also been offering new tools to help marketers reap greater benefits from experimentation. Examples include proper sequential testing (e.g., Pekelis et al. 2015) and multi-armed bandit designs (e.g., Scott 2015, Schwartz et al. 2017).

These concerns and developments in A/B testing mirror developments in academic research (Ioannidis 2005, Head et al. 2015, Open Science Collaboration 2015, Camerer et al. 2018), where difficulties in reproducing research findings have been attributed to two main sources. The first is the preference among many journal editors and research sponsors for novel and statistically significant results. The second is the tendency of authors to engage in behaviors generating significant effects where none exist.

Why would marketers and other business people engage in improper optional stopping or other forms of p-hacking? Unlike academics, they do not face editors biased against publishing null results, nor do they operate under "publish or perish". Hence, they may appear to have little benefit from generating false-positive findings. If anything, rolling out marketing treatments based on false-positive experimental findings would likely hurt rather than boost profits.

Business experimenters may engage in improper optional stopping for the same three reasons as academic researchers do. First, experimenters may simply not have the training or experience to validly interpret statistical test results. Even when experiments are properly conducted, trained practitioners often draw incorrect conclusions from the results. For example, McShane and Gal (2015; 2017) show that not only medical doctors but even highly trained statisticians often make mistakes in interpreting the results of statistical tests. Experimenters may even believe that it is not only defensible but actually best practice to avoid wasting time or subjects and to test for significance repeatedly throughout the experiment without correcting for repeated hypothesis

testing (Miller 2010, John et al. 2012, Johari et al. 2017). Or, they may be aware that optional stopping increases the chances of finding false positives, but deem the procedure a legitimate way to achieve sufficient power without wasting time or subjects because they do not believe that the increase in the false positive rate is disconcertingly large (Simmons et al. 2011, Szucs 2016).

A second possible reason for engaging in improper optional stopping, and the one emphasized by Simmons et al. (2011), is motivated self-deception resulting from the combination of favoring statistical significance and of being uncertain about proper data analysis procedures. As Kareev and Trope (2011) and Yu et al. (2014) note, researchers are often intrinsically motivated to find effects rather than finding none. Sometimes, this preference is in reaction to facing an audience that favors statistical significance. Outside contractors tasked with assessing the effectiveness of a campaign they designed may convince themselves that they are justified to report significant positive results if that helps generate more business. Employees running A/B experiments may be concerned about their perceived competence and prefer to report significant results as well.

A third possible reason why experimenters may engage in improper optional stopping is deliberate deception. Few authors put much weight on this alternative, and several even state explicitly that they deem this factor much less prominent than the other two (Simmons et al. 2011, Yu et al. 2014).

Our study focuses solely on documenting the existence, extent and consequences of p-hacking through improper optional stopping. It does not shed light on whether honest mistakes, self-deception or deliberate deception underlie the behavior.

We investigate to what extent p-hacking occurs in commercial A/B testing, and how it harms the diagnosticity of these tests. Specifically, we investigate to what extent (1) online A/B experimenters stop their experiments based on the p-value of the treatment effect, (2) such improper optional stopping affects the tendency for A/B tests to produce falsely significant results, and (3) consequently depresses the ability of these A/B tests to improve marketing effectiveness.

Our data consists of 2,101 experiments run on the Optimizely A/B testing platform in 2014. The data is unique in that it has daily observations about the performance and termination of each experiment, rather than only the published effect sizes and p-values achieved at termination. Unlike previous analyses using a p-curve (Simonsohn et al. 2014) or a funnel plot of effect sizes vs. standard errors at the time of termination, "looking over the shoulder" of the experimenters in

real-time allows us to directly infer the causal effect of the achieved p-value on stopping behavior through discontinuities within very narrow intervals around critical p-values. Also, our data reflects experimenters' behavior free of selection bias by editors, and focuses solely on improper optional stopping unconfounded by other forms of p-hacking.

Applying a regression discontinuity design (RDD) to the panel data, we estimate that 73% of the experimenters with observations around the 90% confidence level engage in p-hacking. We do not find compelling evidence for p-hacking at 95% or 99% confidence, even though the platform declared "winners" using 95% rather than 90% confidence.

Having documented the existence of improper optional stopping, we proceed to quantifying its consequences on the value of the experimental results. Improper optional stopping is problematic in principle, but how grave the consequences are in practice is an empirical question (Anscombe 1954). For example, the gravity of the problem will depend on the noisiness of the response and the arrival rate of observations between two "peeks" at the data. Our longitudinal data allows us to quantify the consequences of improper optional stopping on the diagnosticity of the tests and their value for improving marketing effectiveness.

Applying the method developed by Storey (2002) and Storey and Tibshirani (2003), we estimate the proportion of all experiments that truly have no effect, regardless of whether the result was declared significant, to be about 73%-77%. This means that the large majority of A/B tests in our sample do not involve treatments of differential effectiveness and hence will not identify more effective business practices. Such a large proportion of true null effects is consistent with the small average treatment effects detected in online advertising experiments (e.g., Lewis and Rao 2015, G. Johnson et al. 2017), but is somewhat lower than the 90% true null rate documented in academic psychological research (V. Johnson et al. 2017). Our estimate that 73%–77% of the effects are truly null supports the suspicion that the high prevalence of non-significant results in A/B tests stems from the interventions being tested rather than the method of A/B testing (Fung 2014).

We estimate that the false discovery rate (FDR) at 90% confidence averages 38% among all experiments, and that p-hacking increases the FDR among p-hackers in our data from 33% to 40%. In other words, p-hacking boosts the probability that an effect declared significant is actually a null effect from 33% to 40%, and by doing so harms the diagnosticity of commercial A/B tests

that use standard null-hypothesis testing. The platform—suspecting that p-hacking was indeed pervasive—deployed advanced sequential testing tools in 2015 to safeguard their users from making such false discoveries even in the presence of p-hacking (Pekelis et al. 2015, Johari et al. 2017).

Since p-hacking increases the average FDR among p-hacked experiments from 33% to 40%, we also quantify the expected cost of a false discovery. Assuming that experimenters stop exploring for more effective treatments after making a false discovery, we estimate the expected cost of a false discovery to be a loss of 1.95% in lift. This corresponds to the 76^{th} percentile of observed lifts.

Finally, our data indicate that several additional experiment characteristics besides statistical significance are associated with stopping behavior. Notably, experimenters facing large negative or large positive effects are more likely to let the experiment continue compared to when observing small effects.

Our study contributes five new insights into how business users stop commercial experiments and how this affects the value of these experiments. (1) We document the existence and quantify the prevalence of optional stopping in commercial A/B testing using data that tracks stopping behavior over time. (2) We estimate the proportion of experiments that truly have no effect. (3) We quantify the impact of stopping behavior on the false discovery rate. (4) We quantify the expected cost of false discovery in terms of forgoing improved lift. (5) We document contingencies in stopping behavior beyond the achieved level of significance.

Our study also makes four contributions to the understanding of p-hacking through optional stopping. (1) It documents the phenomenon among business experimenters rather than academics. (2) It investigates p-hacking through optional stopping specifically whereas other studies cannot distinguish optional stopping from other forms of p-hacking such as adding covariates, adding moderators, analyzing only specific strata of the data, transforming variables, and changing the specific test being used. (3) Using data that "look over the shoulder" of the experimenter throughout the experiment and applying RDD, our work provides stronger causal identification than other work analyzing p-values and effects sizes observed only at the end of the experiment. (4) Our work moves beyond merely documenting the existence and extent of p-hacking, by quantifying its effect on the false discovery rate and providing a method to asses its expected cost in terms of forgoing policy improvements.

2 Data

2.1 Research setting

Our data comes from Optimizely, an online A/B testing platform. It helps experimenters with designing, delivering, monitoring and analyzing different versions of webpages. This section describes the platform as it operated during the data window. An A/B test is a randomized controlled experiment where there are two (A and B) or more versions of a webpage, called webpage variations. When an online user visits the experimenter's website, the platform assigns this visitor to one of the variations, which is then displayed to the visitor. The assignment is usually implemented by saving a cookie file on the visitor's device indicating their assigned variation. Each visitor is assigned to a single variation for the duration of the experiment.

The platform monitors actions that the visitor takes on the website after viewing the assigned variation, and records them in the log of the experiment. The actions being monitored by the platform are chosen by the experimenter and are called "goals". They can include the following:

1. Engagement – How many visitors clicked anywhere on the webpage variation (such as clicking a link or submitting a form)? This is the default goal available on the platform.
2. Click – How many visitors clicked on a specific link or button on the webpage variation?
3. Pageview – How many pageviews or impressions were made on this variation?
4. Revenue – How much sales revenue was generated from this variation?
5. Custom – Other actions defined by the experimenter.

The platform monitors and logs the number of visitors, as well the number of occurrences of the selected actions. The number of occurrences are called the conversion level for each goal. In each experiment, the experimenter designates one variation as the baseline. The baseline may, but need not, be in use before the experiment started. The performance of all other variations is compared to the baseline and statistics are computed relative to the baseline.

The experimenter can access the platform's dashboard and view statistics about the experiment at any time during and after the experiment.[2] The experimenter may terminate the experiment

[2]Throughout the paper, the term experimenter refers to a unique platform account ID, which may be used by

at any time. The platform does not require a pre-set termination time. When the experiment concludes, the experimenter usually picks the variation with the highest performance on the goal of key interest and directs all future visitor traffic to that variation. However, experimenters may also keep the other variations viewable to a small sample of visitors to have a continuously monitored control group.

2.2 Metrics reported to experimenters

The platform displays a dashboard that contains the following metrics for each variation and goal combination in an experiment:

1. Unique visitor and conversion counts. When revenue is the goal, the total revenue rather than the count of conversions is displayed.
2. Conversion rate – the conversion count divided by the unique visitor count.
3. Improvement (in percent) – the relative percent difference in conversion rates between the variation and the baseline. This is also known as the Lift of the variation.
4. Confidence (in percent) – the measure of statistical significance, described below.

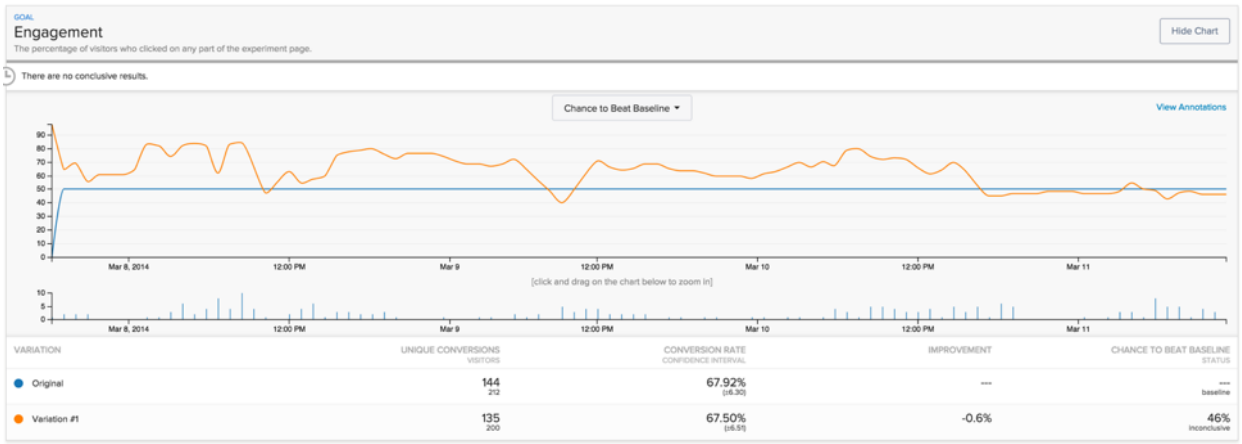
Figure 1 presents such a dashboard. The confidence variable is titled "Chance to beat the baseline" and is calculated as one minus the one-sided p-value of the t-test for the null hypothesis that the conversion rate of the baseline is equal to or greater than that of the variation:

$$H_0 : \mathrm{CR}_{\mathrm{base}} \geq \mathrm{CR}_{\mathrm{var}} \quad \text{vs.} \quad H_1 : \mathrm{CR}_{\mathrm{base}} < \mathrm{CR}_{\mathrm{var}}$$

The confidence value is presented to the experimenter in percentages rounded to two digits (46% in Figure 1). The chart above the statistical metrics in Figure 1 displays how the confidence evolves from the beginning of the experiment until the present. Using two 1-sided tests with an upper and lower threshold, respectively, is equivalent to using one 2-sided test with a single threshold. In the case just described, the two 1-sided tests using 95%/5% confidence are equivalent to a single 2-sided test with 10% significance or 90% confidence.

multiple individuals. Hence we use the term experimenter to denote either a unique individual or a set of individuals running A/B tests using the same account ID.

Figure 1: Experimenter Dashboard: Overview



When the confidence value of a variation reaches a pre-determined upper threshold (95% is the default), that variation is displayed as a "winner". When its confidence reaches a lower threshold (5% is the default), the variation is displayed as a "loser". When the confidence value is between the two thresholds, the result is displayed as inconclusive. See Figures 2a, 2b and 2c for examples. Note, since the default notification is set at 5% and 95% confidence, any p-hacking at other levels cannot be attributed to the default notification by the platform.

2.3 Set of experiments studied

Our raw data contains the entire set of 9,214 experiments that were started on the platform during the month of April 2014. All but one ended by November 30 2014, the end of our observation window. The data contains daily values of visitor and conversion counts for each variation in each experiment, from which we calculate the metrics and statistics used in the analysis.

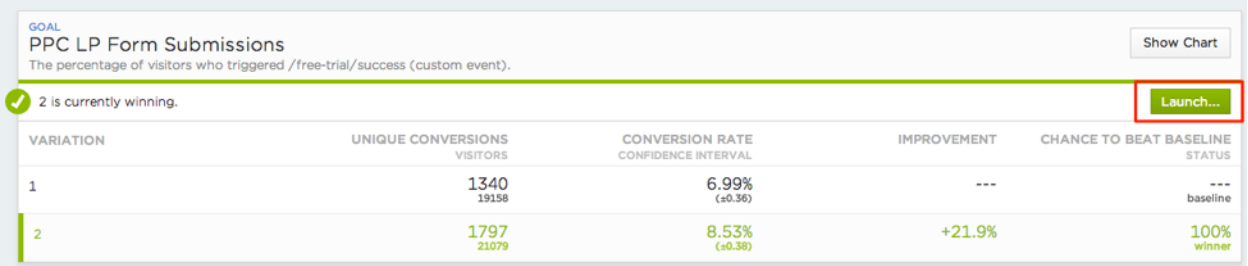
Since we seek to cleanly identify p-hacking in termination behavior, we exclude experiments that have one or more of the following characteristics precluding or harming clean identification:

1. Having exactly the same number of visitors and converters in all variations at all times.
2. Having all traffic occur on only a single variation for six or more consecutive days.
3. Having no traffic to any variation for six consecutive days.

Experiments with the first characteristic are most likely pre-tests where exactly the same web-page appears in all variations. Such pre-tests are recommended by platforms. Experiments with

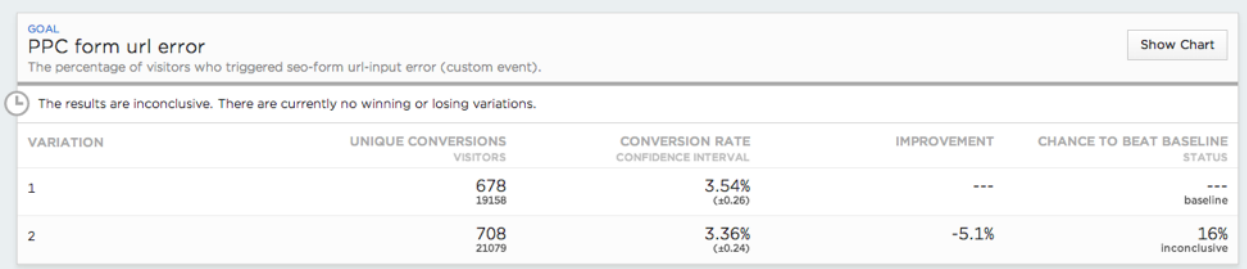
Figure 2: Experimenter Dashboard: Winning, Losing and Inconclusive Variations

(a) Winning Variation



(b) Losing Variation

(c) Inconclusive Variation



the second or third characteristics very likely were terminated de facto by reconfiguring the website before the experimenter notified the platform about the experiment's termination.

We further purify the data by including only experiments for which the industry of the experimenter is reported and for which Engagement is listed as a goal. The latter allows us to compare performance metrics on the same goal across experiments. Because Engagement is the default goal and the most popular goal, focusing on Engagement also results in the largest set of experiments for us to study. Only 12% of the experiments list Engagement as the only goal, and it is possible that some experimenters pursued a primary goal other than Engagement. We do not have that information.

Finally, 34% of the experiments had two or more non-baseline variations. For such cases, we select the variation with the highest conversion rate on the last day of the experiment as the

primary variation. This allows us to compare experiments with different numbers of non-baseline variations. We choose the variation with the highest conversion rate under the assumption that the experimenter intends to deploy the best performing treatment.

Our final dataset consists of 2,101 experiments from 916 experimenter accounts. The day of termination is observed for all these experiments. The data contains 76,215 experiment-day observations. For each experiment-day we observe the number of visitors (sample size) to each variation, the number of clicks (Engagement) in each variation, and whether the experiment was terminated on that day.

The experiments come from various industries (Table 1). Retailers and E-tailers (26.5%), High-tech companies (17.7%), Media companies (15.8%) and Professional Services providers (7.8%), account for the majority of experiments. Table 2 reports additional characteristics of the experiments. The median number of variations including the baseline was 2, and the median number of goals was 3. Experimenters varied quite a bit in the number of prior experiments they had run on the platform, with the median being 184. On the last day, the typical (median) experiment included more than 10,000 visitors and had run for 19 days.

Table 1: Distribution of Experiments by Industry

Industry Vertical	Percentage
Financial Services	2.09
Gaming	0.14
High Tech	17.71
Insurance	0.33
Media	15.75
Mobile Only	0.05
Non-Profit	1.14
Other	17.99
Professional Services	7.81
Retail	26.56
Schools & Education	3.00
Telecommunications	1.14
Travel & Entertainment	6.28

Industry assigned by the platform. N=2,101.

Table 2 also reports two performance metrics on the day the experiment ended. The Effect Size of a non-baseline variation is the difference in conversion rates between that variation and the baseline, whereas the Lift is the percentage difference in conversion rates from the baseline. Lift is reported as "Improvement" on the dashboard (Figure 1). As noted earlier, we select the non-baseline variation with the highest conversion rate on the last day to characterize the time-varying

Table 2: Summary Statistics of Experiments

	Mean	Median	SD	Min	Max
Total Variations	2.78	2.00	2.66	2.00	80.00
Total Goals	4.67	3.00	5.47	1.00	70.00
Past # Experiments	364.3	184.0	455.7	1.0	2,917.0
Sample Size	141,309	10,129	915,721	201	34,724,196
Length (in Days)	36.28	19.00	49.19	1.00	537.00
Effect Size	0.005	0.001	0.047	−0.328	0.571
Lift	0.112	0.0015	2.256	−0.653	82.78

$N = 2,101$. Values are computed on the last day of the experiment.

Effect Size, Lift and Confidence values of the experiment. When the lift is undefined on a particular experiment-day because of a zero baseline conversion rate, we set it to zero if the effect size was zero, and to the highest 99% percentile if the effect size was positive. This affects 0.13% and 0.19% of experiment-days, respectively. As reported in Table 2, the average Effect Size was 0.5% and the average Lift was 11.2%, meaning that the Engagement in the best-performing variation was on average half a percentage point or 11.2% higher than in the baseline.

3 When do experiments end?

3.1 Model Free Evidence

The propensity to start or end experiments varies across the days of the week. As one might expect, few experiments start or end during the weekend (Table 3). However, the propensity to start experiments is also markedly lower on Mondays and Fridays compared to the rest of the week, as is the propensity to end on Fridays.

Table 3: Days of the Week when Experiments Start and End (%)

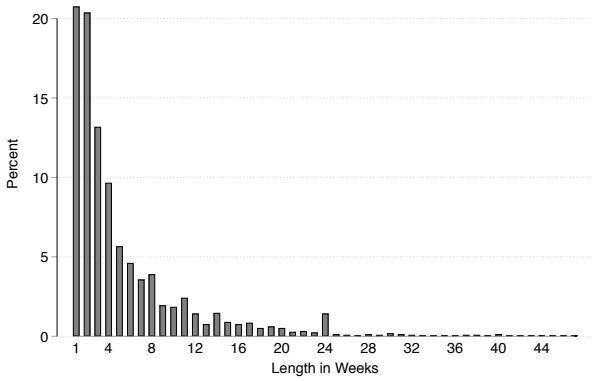
	Start	End
Monday	11.14	21.89
Tuesday	20.32	19.70
Wednesday	22.13	17.85
Thursday	17.71	20.18
Friday	14.85	15.42
Saturday	7.33	2.52
Sunday	6.52	2.43
Total	100.00	100.00

$N = 2,101$.

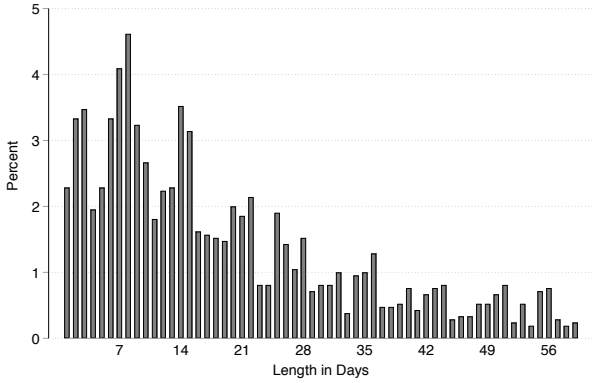
Figure 3a shows the histogram of how long experiments run. Half the experiments end in less than three weeks. Specifically, 25% end in 8 days or less, 50% in 19 days or less, 75% in 43 days or less, and only 5% last longer than 132 days. A small spike in the number of endings between 161 and 168 days is notable, reflecting a tendency not to run experiments beyond 24 weeks. Figure 3b shows the histogram of durations up to 60 days, which covers 82.7% of all experiments. Clearly, there is a downward trend and a 7-day cycle. The cycle may stem from the tendency to start and end the experiment on particular weekdays (Table 3), but may also reflect a tendency to run experiments in multiples of weeks.[3]

Figure 3: Histograms of Experiments' Length

(a) All days



(b) First 60 days



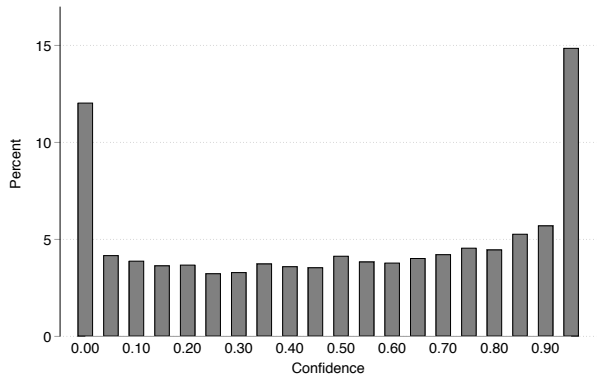
Ending an experiment is also associated with the level of confidence achieved. Figure 4 presents histograms of confidence values on any day during the experiment (4a) and on the day they end (4b). There is slightly more mass at confidence levels of values .85 and higher in Figure 4b than Figure 4a. This slight skew might, but need not, stem from p-hacking, since statistical power increases over time as the sample size grows. The distribution of lift shows a similar slight skew towards high values on the last day (Figures 5a and 5b). However, the dominant feature of both histograms is their peakedness around zero.

Figure 6 plots the empirical hazard of stopping the experiment on a particular day against the confidence level achieved on that day, binned in 10% increments. The slope of the grey line indicates that when the confidence level is below 0.5 and hence the lift is negative, there is a very

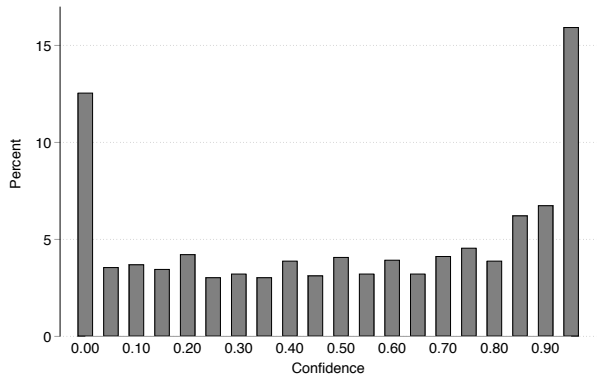
[3]The empirical hazard rate—the number of experiments ending on a day divided by the number of experiments that did not end earlier—exhibits the same patterns: a downward trend for about 270 days (by which time 99.2% of the experiments have ended), a 7-day cycle in the first 60 days, and a spike in week 24.

Figure 4: Histograms of Experiments' Confidence

(a) All days



(b) Last day of experiments



slight negative relation between confidence level and the hazard of stopping. In contrast, when the confidence level is above 0.5 and hence the lift is positive, there is a slight tendency to end the experiment when the confidence level is higher.

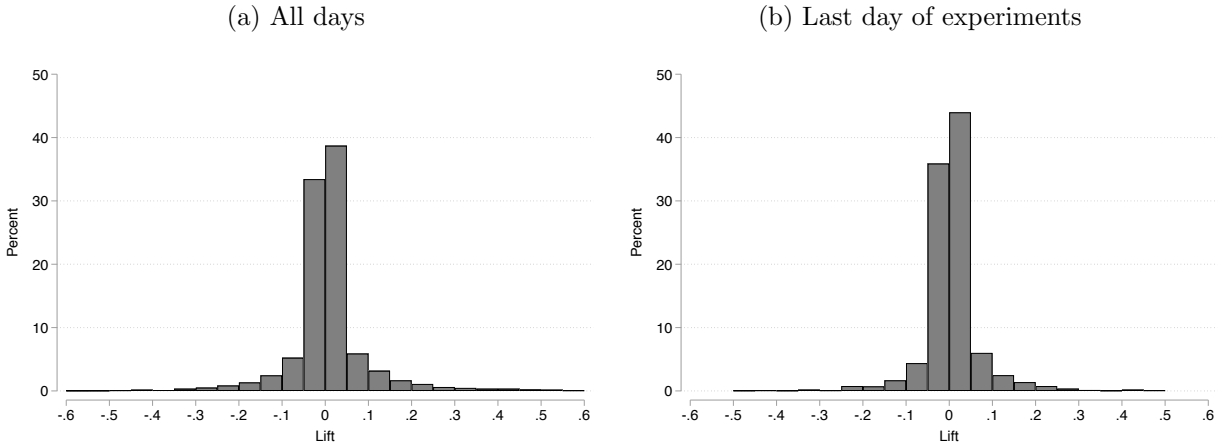
The orange and blue line segments in Figure 6 indicate that the hazard of stopping on a particular day is also associated with lift, and that this association depends on whether the lift is positive or negative. When the confidence level is below 0.5 and hence the lift is negative, higher (less negative) lifts are associated with stopping *sooner*. In contrast, when the confidence level is above 0.5 and hence the lift is positive, higher (more positive) lifts are associated with stopping *later*. In short, experimenters observing very negative or very positive effect sizes are more likely to let the experiment continue to run compared to when observing small effects.

3.2 Hazard Model Analysis

The model free analysis shows that stopping experiments on a particular day is associated with how many days they have been running, the day of the week, the effect size, and the confidence level or p-value achieved. The association between stopping and the p-value may stem from experimenters being exceedingly good at setting the pre-determined sample size or run-time to detect the experimental effect, but might also reflect p-hacking. Hazard modeling allows us to incorporate all these predictors in a single analysis of stopping behavior.

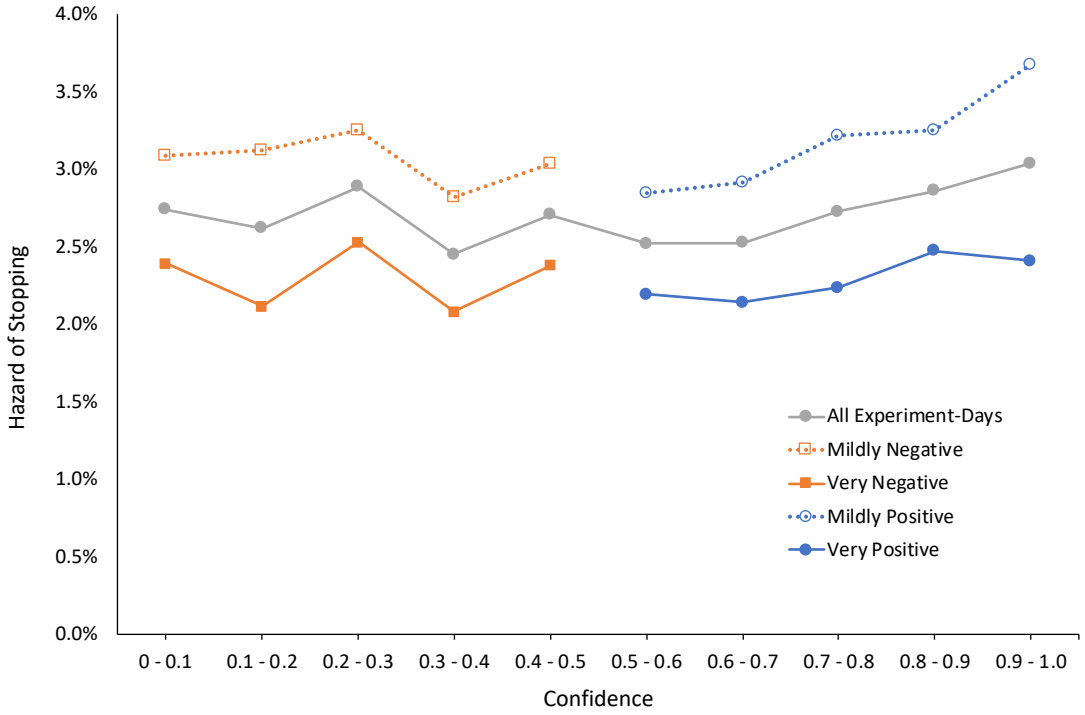
Since the data are observed daily and predictors vary over time, we use a discrete time hazard modeling approach. We include experimenter specific random intercepts to account for possible

Figure 5: Histograms of Experiments' Lift



Lift values between -0.02 and 0.02 constitute 47.0% (left) and 71.5% (right) of experiment-days.
Lift values between -0.60 and 0.60 constitute 97.3% (left) and 99.3% (right) of experiment-days.

Figure 6: Empirical Hazard of Stopping the Experiment by Confidence and Lift



correlation in stopping behavior of experiments managed by the same experimenter.

Let y_{ijt} equal 1 when experimenter i terminates experiment j after t days, and equal 0 otherwise. We model the hazard $h_{ijt} = Pr(y_{ijt} = 1|y_{ijt-1} = 0) = F(\alpha_i + x_{ijt}\beta)$, where $F(\cdot)$ is the logistic cdf, $\alpha_i \sim \mathcal{N}(0, \sigma^2)$, and x_{ijt} are observed covariates.

The two variables of main substantive interest are confidence and lift. Mirroring Figure 6, we bin confidence levels in 10% increments, and effects-code lift as $+1/-1$ reflecting whether the value is above or below the bin-specific median.

In addition, we include the following controls:

1. Cumulative number of days the experiment has been running so far (duration dependency). We use a piece-wise constant baseline hazard specification. The hazard varies freely from day to day between days 1 and day 35, after which it changes freely in 5 day intervals up to day 100, and 50 day intervals up to day 250.

2. Day of the week. We use a separate dummy for each day of the week.

3. Cumulative number of visitors, i.e., the time-varying sample size of the experiment.

4. Industry. We use a dummy for each industry reported in Table 1.

5. Past # Experiments. This controls for experimenters' experience in using the platform.

Table 4 presents the estimates of two models. The first model quantifies, for each confidence bin, the propensity (log-odds) of stopping and the extent to which it varies between experiment-days above or below the median lift in that bin. The second does so as well, but includes the entire set of control variables.

The coefficients in the first ten rows capture the relation between confidence and stopping at the median lift. Though the coefficients exhibit a weak U-shape in both models, consistent with Figure 6, omnibus Wald tests indicate that the coefficients are not significantly different from each other in either model ($p > 0.10$). The coefficients in the next five rows of Model (1) indicate that when confidence $< 50\%$, a higher (less negative) lift is associated with stopping earlier. The pattern in Model (2) including the control variables is similar but less pronounced. The next five rows of coefficients in Models (1) and (2) indicate that when confidence $\geq 50\%$, a higher (more positive) lift is associated with stopping later rather than earlier. This reversal in how lift moderates the

association of confidence with stopping reflects the pattern observed in Figure 6. Table A1 in the Online Appendix reports the full set of coefficients, including those of the control variables. It is worth noting that the 7-day cycle and the day of the week association found in the model-free analysis are also present after controlling for confidence, lift and sample size.

Table A1 also presents estimates of models where lift is not coded as $+1/-1$ to indicate being above/below the bin-specific median, but where lift is coded as the actual or normalized lift centered around the bin-specific mean. The sign reversal of the interaction between confidence and lift is robust.

4 Do experimenters p-hack?

The model-free and model-based results reported so far do not establish causality. Consequently, they do not document p-hacking through optional stopping. To what extent do people stop the experiment *because* they have reached a critical confidence level?

Even though our data are not experimental, we can achieve causal identification by exploiting the panel structure and the fine granularity of our data within a regression discontinuity design. Specifically, we assess whether people p-hack by investigating whether their propensity to stop their experiments "jumps" at critical confidence levels.

We test six hypotheses. Each asks whether experimenters p-hack at one specific level. These are 90%, 95% and 99% for positive effects, and 1%, 5% and 10%, which are the equivalent levels for negative effects. We focus on these levels because they correspond to the levels of significance commonly used for hypothesis testing (Leahey 2005, Brodeur et al. 2018).

Because we have access to the actual confidence values, but users of the platforms only see values rounded to two digits, the actual critical levels we use are: .015, .055, .105, .895, .945 and .985. For example, to assess whether people p-hack when the dashboard reports 90% confidence, we use the .895 threshold.

In small windows around these thresholds, the confidence achieved by the experiment is practically random. The reason is that confidence, i.e., a t-statistic converted to a probability, is a function of effect size and sample size, both of which are driven by the random process of visitor arrival to the site.

Table 4: Hazard Regression Results for Confidence and Lift

	(1)	(2)
Conf. 0.0–0.1	-3.3453^{***}	-4.3611^{***}
	(0.0000)	(0.0000)
Conf. 0.1–0.2	-3.4056^{***}	-4.3498^{***}
	(0.0000)	(0.0000)
Conf. 0.2–0.3	-3.3292^{***}	-4.2563^{***}
	(0.0000)	(0.0000)
Conf. 0.3–0.4	-3.4927^{***}	-4.4421^{***}
	(0.0000)	(0.0000)
Conf. 0.4–0.5	-3.4212^{***}	-4.3922^{***}
	(0.0000)	(0.0000)
Conf. 0.5–0.6	-3.5222^{***}	-4.4606^{***}
	(0.0000)	(0.0000)
Conf. 0.6–0.7	-3.4880^{***}	-4.4327^{***}
	(0.0000)	(0.0000)
Conf. 0.7–0.8	-3.3768^{***}	-4.3384^{***}
	(0.0000)	(0.0000)
Conf. 0.8–0.9	-3.3572^{***}	-4.3303^{***}
	(0.0000)	(0.0000)
Conf. 0.9–1.0	-3.3230^{***}	-4.3195^{***}
	(0.0000)	(0.0000)
Conf. 0.0–0.1× lift	0.1168^{*}	0.0304
	(0.0638)	(0.6630)
Conf. 0.1–0.2× lift	0.2562^{***}	0.1730^{*}
	(0.0041)	(0.0658)
Conf. 0.2–0.3× lift	0.1928^{**}	0.1035
	(0.0274)	(0.2586)
Conf. 0.3–0.4× lift	0.1966^{**}	0.1074
	(0.0359)	(0.2733)
Conf. 0.4–0.5× lift	0.2285^{***}	0.1845^{**}
	(0.0095)	(0.0449)
Conf. 0.5–0.6× lift	-0.1584^{*}	−0.1113
	(0.0645)	(0.2128)
Conf. 0.6–0.7× lift	-0.2322^{***}	-0.1652^{*}
	(0.0081)	(0.0720)
Conf. 0.7–0.8× lift	-0.2860^{***}	-0.2137^{**}
	(0.0004)	(0.0125)
Conf. 0.8–0.9× lift	-0.2148^{***}	-0.1559^{**}
	(0.0043)	(0.0497)
Conf. 0.9–1.0× lift	-0.1924^{***}	-0.1304^{**}
	(0.0004)	(0.0298)
Sample Size		−0.0000
		(0.2141)
Past # Experiments		−0.0001
		(0.2121)
Industry FE	No	Yes
Day of Week FE	No	Yes
Day FE	No	Yes
LL	−9257.273	−8696.030
σ	0.752	1.151

N = 76,215. # of Experimenters = 916.

p-values in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

4.1 RDD Analysis

We follow the continuity approach in regression discontinuity designs (Imbens and Lemieux 2008). Our approach is somewhat different from the typical RDD application because of the panel-data and hazard setting (Bor et al. 2014).[4] Consequently, we can observe the same experiment both below and above the cutoff.

We consider each of the six critical confidence levels as a suspected discontinuity point, set the bandwidth of the window around each critical level, and then estimate the size and significance of the discontinuity within the window. Each data window is symmetric and its width is twice its bandwidth. Choosing between wide and narrow windows involves a trade-off. Wider windows contain more observations and hence typically yield higher power, whereas narrower windows typically offer lower bias. One popular method to trade off bias against precision is to use the bandwidth that minimizes the mean square error (MSE) of the discontinuity estimate (Imbens and Kalyanaraman 2012, Calonico et al. 2014). We use this bandwidth selection method and combine it with a triangular kernel, which results in a point estimator with optimal properties in terms of MSE.

Next we estimate two local linear regressions, one on each side of the critical level. Each includes a linear and a quadratic term for confidence level. Following Calonico et al. (2014), Gelman and Zelizer (2015), and Gelman and Imbens (2018), we do not consider higher order polynomials. We estimate the discontinuity from the local linear regressions, using both conventional and bias-corrected estimators (Calonico et al. 2014) and cluster-robust standard errors for inference. Optimal bandwidth selection, local linear regression and inference are conducted using the rdrobust package (Calonico et al. 2017).

Table 5 describes the inputs and outputs of the bandwidth selection procedure. The first row reports the data ranges (expressed in confidence levels) used as raw data for each critical threshold. They are specified such that they are non-overlapping and symmetric. The second row reports the number of observations in the data range. The third row reports the MSE-optimal bandwidth. Rows four and five report the number of observations to the left and the right of the critical threshold within the MSE-optimal window.

The local linear regressions estimated using the MSE-optimal bandwidths are linear probability

[4]The RDD hazard analysis exploits the "looking over the shoulder" panel structure of the data and the operationalization of p-hacking in terms of stopping time. This is unlike funnel plots and p-curves used in analyses of p-hacking in academic settings, that analyze only terminal effect sizes, standard errors and p-values.

Table 5: Regression discontinuity windows

Cutoff	.015	.055	.105	.895	.945	.985
Data Range	(0, 0.03)	(0.03, 0.08)	(0.08, 0.13)	(0.87, 0.92)	(0.92, 0.97)	(0.97, 1)
N	6,704	3,390	2,936	4,093	4,669	6,222
Optimal Bandwidth	0.0040	0.0109	0.0081	0.0113	0.0075	0.0057
Eff. N Left	346	732	482	970	648	618
Eff. N Right	300	775	471	924	658	751

models, which may produce biased estimates for rare events (Horrace and Oaxaca 2006). We therefore complement these analyses with a logit hazard model, using the same bandwidths and a uniform kernel. The logit specification for the hazard that i stops experiment j on day t is:

$$h_{ijt} = F(\alpha_i + \beta_D \cdot D_{ijt} + \beta_X \cdot X_{ijt} + \beta_{DX} \cdot D_{ijt} \cdot X_{ijt} + \beta_{X^2_{ijt}} \cdot X^2_{ijt} + \beta_{DX^2} \cdot D_{ijt} \cdot X^2_{ijt})$$

where $F(\cdot)$ is the logistic cdf, $\alpha_i \sim \mathcal{N}(0, \sigma^2)$, X is confidence minus the critical level, and D indicates whether the confidence is above or below the critical level.[5]

In this model β_D captures the size of the discontinuity in the log-odds of stopping at the critical threshold. As such, a significant value of β_D represents the causal effect of reaching a p-value on stopping. For p-hacking to exist, the effect has to be positive for positive treatment effects (90%, 95%, 99%) and negative for negative treatment effects (10%, 5%, 1%). To facilitate comparison with the linear model estimates, we report the estimated marginal effects rather than β_D.

4.2 RDD Results

We test six hypotheses, each asking whether experimenters p-hack at a specific critical threshold. Table 6 presents the results. For each threshold it reports three estimates of the jump in probability of terminating the experiment at the threshold. First, the estimate of the local liner regression; Second, that of the local linear regression with bias correction; Third, that of the logistic regression.

We find compelling evidence of a discontinuity only at the 90% confidence threshold (.895 cutoff). All three estimates of the discontinuity are significant ($p < 0.05$). The probability of stopping jumps up by roughly 7 percentage points, from 1% to 8%.

[5]The RDD analysis excludes observations for experiments terminated before reaching the confidence window on any day. Consequently, effect sizes are computed only for the population of experiments falling in the windows, but the estimates do not suffer from truncation bias (Van den Bulte and Iyengar 2011).

There is some weak evidence of a discontinuity at the 95% confidence threshold (.945 cutoff), but with only one of the three estimates being significant ($p < 0.05$), it is far from compelling. There is no evidence of discontinuities at any of the four other levels. Hence we reject the null hypothesis of no regression discontinuity only for the 90% confidence threshold.

In summary, people p-hack only for positive effects and only at the most liberal level of 90% confidence. Assuming that the baseline is the status quo, such that a positive lift is a success and a negative lift is a failure, experimenters "pull the p-hacking trigger" when facing good news rather than bad news. Importantly, the 90% confidence level at which they p-hack is *not* the level highlighted by the platform which calls a variation a winner when the confidence level reaches 95%, and a loser when the confidence level reaches 5%. Consequently the evidence of p-hacking *cannot* be attributed to visual salience or to the declaration of a winner or loser by the platform.[6]

Table 6: Regression discontinuity estimates

	Cutoff .105			Cutoff .895		
	Effect	p-value	95% C.I.	Effect	p-value	95% C.I.
Linear	−0.0129	0.6160	[−0.0635, 0.0376]	0.0615**	0.0415	[0.0024, 0.1205]
Linear, Bias Corr.	−0.0193	0.4810	[−0.0731, 0.0345]	0.0714**	0.0297	[0.0070, 0.1357]
Logistic, R.E.	−0.0116	0.9988	[−15.7545, 15.7313]	0.0777**	0.0300	[0.0075, 0.1478]
	Cutoff .055			Cutoff .945		
	Effect	p-value	95% C.I.	Effect	p-value	95% C.I.
Linear	0.0109	0.6395	[−0.0347, 0.0565]	0.0643*	0.0538	[−0.0010, 0.1297]
Linear, Bias Corr.	0.0073	0.7779	[−0.0432, 0.0578]	0.0750**	0.0431	[0.0023, 0.1476]
Logistic, R.E.	0.0118	0.6418	[−0.0378, 0.0613]	0.0488	0.1783	[−0.0222, 0.1198]
	Cutoff .015			Cutoff .985		
	Effect	p-value	95% C.I.	Effect	p-value	95% C.I.
Linear	0.0027	0.8487	[−0.0247, 0.0301]	−0.0407	0.1985	[−0.1028, 0.0214]
Linear, Bias Corr.	0.0035	0.8209	[−0.0266, 0.0336]	−0.0503	0.1612	[−0.1207, 0.0201]
Logistic, R.E.	−0.0072	0.5477	[−0.0306, 0.0162]	−0.0246	0.9918	[−4.7209, 4.6717]

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

4.3 RDD Validation Checks

Several checks validate the presence of a discontinuity at the 90% threshold.

First, we conduct placebo tests investigating the presence of a discontinuity at the 89% and

[6]Users were able to instruct the platform to declare winners and losers at 90%/10%, but this override of the default paired with a jump in the probability of stopping at 90% would constitute another indication of p-hacking. Moreover, we do not find evidence of p-hacking at 10% which makes it even less credible that mere visual salience in experiments with this override is the mechanism driving the discontinuity at 90%.

91% confidence levels. We repeat the bandwidth selection and discontinuity estimation procedures outlined above, but now for cutoffs at .885 and .905. We find no evidence of any discontinuities at either level (Table 7).

Table 7: Placebo tests of regression discontinuities at .885 and .905 compared to .895 (reported to the user as 89%, 91% and 90%)

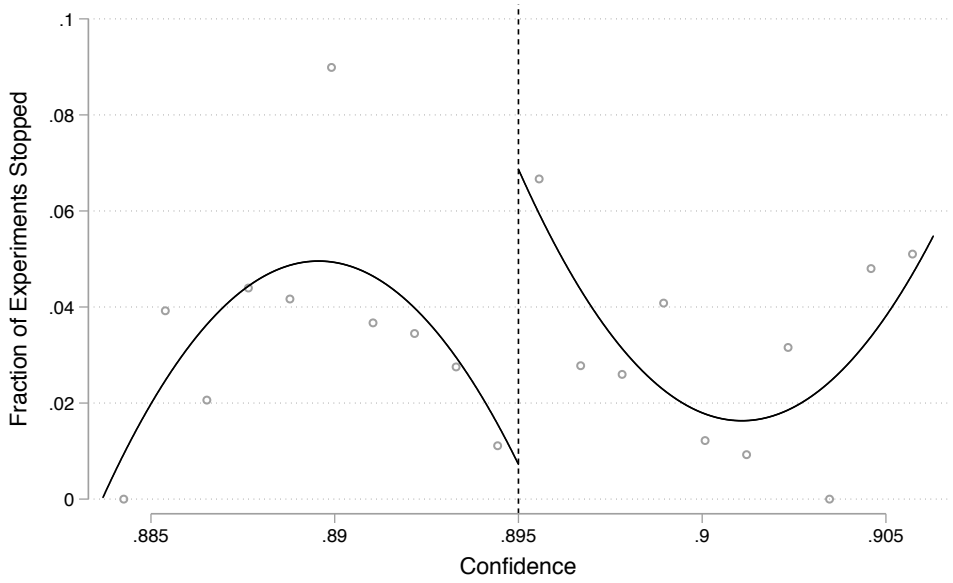
	Cutoff .885		
	Effect	p-value	95% C.I.
Linear	−0.0169	0.5467	[−0.0720, 0.0381]
Linear, Bias Corr.	−0.0252	0.4263	[−0.0874, 0.0369]
Logistic, R.E.	0.0000	0.9982	[−0.0375, 0.0376]
	Cutoff .905		
	Effect	p-value	95% C.I.
Linear	−0.0150	0.5881	[−0.0692, 0.0392]
Linear, Bias Corr.	−0.0199	0.5255	[−0.0813, 0.0415]
Logistic, R.E.	−0.0129	0.6691	[−0.0723, 0.0464]
	Cutoff .895		
	Effect	p-value	95% C.I.
Linear	0.0615**	0.0415	[0.0024, 0.1205]
Linear, Bias Corr.	0.0714**	0.0297	[0.0070, 0.1357]
Logistic, R.E.	0.0777**	0.0300	[0.0075, 0.1478]

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Next, as recommended by Gelman and Zelizer (2015), we present a discontinuity plot showing both the data and the fitted model (Figure 7). We organize the data in 20 bins of confidence and then compute the empirical hazard rate within each bin, i.e., the fraction of experiment-days on which an experiment is stopped. Circles indicate the empirical hazard. A circle with zero value indicates no stops within that bin. Lines indicate the predicted value of the local linear regression models. The presence of the discontinuity is not an artefact of the quadratic local linear regression, but is supported by the raw data as indicated by the circles just to the left and just to the right of .895.

Third, we check for the absence of discontinuities of four other variables associated with stopping behavior (Section 3): number of days the experiment has been running, number of visitors (sample size), whether it was a weekend or not, and the lift. Figure A1 presents histograms of these covariates binned by the level of confidence. There are no pronounced jumps at the .895 critical

Figure 7: Regression discontinuity plot of stopping at .895 (Shown to user as 90%)



level or any trends throughout the window. This is consistent with the key assumption for causal identification that the continuous confidence variable behaves randomly within the narrow window around the critical level. Table A2 reports RDD analyses for those four variables, and shows no evidence of discontinuity for any of these four variables at the .895 confidence threshold.

4.4 What fraction of experimenters p-hack at 90% confidence?

The discontinuity documented so far pertains to the *average* experiment-day within the specified windows. Another question of interest is what fraction of experimenters p-hack once their experiments have reached the window. We address this question for the 90% critical level because that is the only level where we find compelling evidence of p-hacking (Table 6).

We estimate a latent class logit hazard model where experimenters either p-hack or not. A p-hacker is someone whose stopping behavior is affected by reaching 90% confidence, whereas a non p-hacker's stopping behavior is not. We allow the stopping behavior of p-hackers to exhibit discontinuities and specify the logit-hazard of a p-hacker v^P_{ijt} as:

$$v^P_{ijt} = \beta_0 + \beta_X \cdot X_{ijt} + \beta_{X^2} \cdot X^2_{ijt} + \beta_D \cdot D_{ijt} + \beta_{DX} \cdot D_{ijt} \cdot X_{ijt} + \beta_{DX^2} \cdot D_{ijt} \cdot X^2_{ijt}$$

For non p-hackers, we do not allow discontinuities and impose $\beta_D = 0$, $\beta_{DX} = 0$ and $\beta_{DX^2} = 0$:

$$v_{ijt}^{NP} = \beta_0^{NP} + \beta_X^{NP} \cdot X_{ijt} + \beta_{X^2}^{NP} \cdot X_{ijt}^2$$

We use the same bandwidth as in the RDD analysis. This results in $1,894$ experiment-days from 379 of the original 916 experimenters.

The likelihood of the data, maximized using expectation-maximization (EM), is:

$$\prod_{i=1}^{n} \prod_{j=1}^{J_i} \prod_{t=1}^{T_{ij}} \left(wF(v_{ijt}^P)^{y_{ijt}} (1 - F(v_{ijt}^P))^{(1-y_{ijt})} + (1-w)F(v_{ijt}^{NP})^{y_{ijt}} (1 - F(v_{ijt}^{NP}))^{(1-y_{ijt})} \right) \quad (1)$$

where i indicates the experimenter, j indicates the experiment, t indicates the experiment-day, n is the number of experimenters, y_{ijt} indicates whether the experiment was stopped, and w is the proportion of p-hackers.

We estimate the discontinuity β_D to be 1.85 [95% CI: 0.19, 3.51], and the fraction of p-hackers w to be 72.65% [95% CI: 53.21%, 86.13%].[7] Hence, when experiments are within the $89.5\% \pm 1.13\%$ confidence window, roughly three quarters of the experimenters p-hack. The estimates of the intercept and β_D correspond to a jump in the probability of stopping from 0.0222 to 0.1259 which amounts to a marginal effect of 0.1038 for p-hackers. Multiplying this by the fraction of p-hackers (72.65%) reproduces the logit model average marginal effect from Table 6, 0.0777.

We also estimate a variant of the model where we asses whether the experience or the industry of the experimenter are associated with being a p-hacker. We find no statistically significant patterns (Table A3 in the Online Appendix).

5 What fraction of significant results are false positives?

p-Hackers reject the null hypothesis too often. Specifically, when people p-hack at 90%, they reject true null effects more than 10% of the time. In terms of Table 8, given m_0 cases with a true null effect, p-hackers inflate the number of cases that is called significant, F, and deflate the number called not-significant, $m_0 - F$. By inflating F, p-hackers also end up inflating the fraction of significant results that are truly null. Inflating the false positive rate $FPR = \frac{F}{m_0}$ also inflates the

[7] Table A3 reports the values for all coefficients.

false discovery rate $FDR = \frac{F}{S}$. Consequently, experimenters and their audience end up believing that a greater number of treatments or interventions are effective than is warranted.

Table 8: False Positives and False Discoveries

	Called Significant (Discovery)	Called not Significant	Total
Null is True	F	$m_0 - F$	m_0
Alternative is True	T	$m_1 - T$	m_1
Total	S	$m - S$	m

In this Section we quantify by how much p-hacking inflates the FDR in the experiments we study. We start by estimating the fraction of true nulls $\pi_0 = \frac{m_0}{m}$. Although the treatment effect on continuous conversion rates is unlikely to be exactly zero, allowing for the existence of true nulls ($\pi_0 > 0$) is consistent with the notion that null effects include negligibly small effects, centered at 0, that require extremely large samples to detect (Hodges and Lehmann 1954, Berger and Delampady 1987, Masson 2011).

We estimate π_0 and the expected FDR, $\mathbb{E}\left(\frac{F}{S}\right)$, using the method developed by Storey (2002) and Storey and Tibshirani (2003) who show that when the number of hypotheses m is large, $\mathbb{E}\left(\frac{F}{S}\right) \approx \frac{\mathbb{E}(F)}{\mathbb{E}(S)}$. $\mathbb{E}(S)$ is the expected number of rejected hypotheses, which is estimated simply as the number of significant p-values on the last day of the experiment. $\mathbb{E}(F)$ is estimated as $m \cdot p \cdot \pi_0$ where m is the number of experiments in the data, p is the 2-sided nominal significance level, and π_0 is estimated using the Storey and Tibshirani (2003) method.[8] We perform this analysis on 2,053 experiments with at least two days per experiment, because a subsequent analysis requires observations on both the last and the preceding day, and hence precludes using experiments lasting only one day. We use bootstrapping with 3,000 samples to compute bias-corrected and accelerated (BC_a) confidence intervals for the estimated FDRs.

Table 9 presents the estimates of π_0 and the implied FDRs for the 90%/10%, 95%/5% and 99%/1% discovery thresholds. There are two main insights. First, about 73% of the experiments are estimated to have true null effects. In other words, when performing A/B tests about 3/4 of experiments should not have been expected to yield an improvement over the baseline. Second, 38% of the results significant at 90%/10% are true nulls, as are 27% of those significant at 95%/5%, and 11% of those significant at 99%/1%.

[8] When applying the method of Storey and Tibshirani (2003), we account for the fact that their formulas assume 2-sided tests, whereas the p-values computed in our experiments are for 1-sided tests.

Table 9: Fraction of True Null Effects and FDRs

	Estimate	95% C.I.
π_0	72.67%	[62.83%, 84.12%]
FDR 90% / 10%	38.07%	[32.27%, 45.00%]
FDR 95% / 5%	26.89%	[22.46%, 31.60%]
FDR 99% / 1%	10.52%	[8.86%, 12.91%]

Having estimated the FDRs in the experiments, we now proceed with quantifying the magnitude by which p-hacking inflates those FDRs. To perform this analysis, we compare the actual FDR, i.e., based on p-values on the *last day*, to what the FDR would have been if the experiment had ended the day before. That day before termination serves as the counterfactual, since it is the nearest-neighbor of the day of termination, but for which we know there was no p-value based termination. We compute that counterfactual FDR in two ways: (1) based on the p-values the day before the experiment was stopped (FDR_1^{CF}), and (2) using the effect size from the day before the experiment was stopped, but the sample size from the last day (FDR_2^{CF}). Because p-values are random around the stopping threshold, both counterfactual FDRs reflect the FDR without p-hacking. However, comparing the actual FDR to FDR_1^{CF} does not control for the larger sample size observed on the last day which expectedly will result in a greater test statistic and hence deflate the estimated effect of p-hacking. Using the second countefactual FDR_2^{CF} corrects for this difference in sample size.

To estimate the causal impact of p-hacking at 90%/10% confidence on the FDR, we exploit the fact that false discoveries come from a mixture of p-hackers (P) and non p-hackers (NP). The actual FDR on the last day equals:

$$FDR = wFDR^P + (1-w)FDR^{NP} \tag{2}$$

where w is the fraction of p-hackers which we estimated to be 72.65% (Section 4.4).[9] The FDR if

[9]When using $w = 72.65\%$, we assume that the fraction of p-hackers at 90%/10% in the full sample of 916 experimenters is the same as the fraction of p-hackers among the 379 experimenters with experiments in the $89.5\% \pm 1.13\%$ confidence window. Since we did not find evidence of p-hacking at 10%, this assumption likely leads to underestimating the quantity in Equation (4) and the impact of p-hacking on the FDR of p-hackers.

none of the experimenters had p-hacked, FDR_i^{NP}, is measured using the counterfactuals:

$$FDR_i^{NP} = FDR_i^{CF} \text{ where } i = 1, 2 \tag{3}$$

Consequently, the estimated impact of p-hacking on the FDR of p-hackers, similar to an estimate of the average treatment effect on the treated (ATT), equals:

$$\Delta FDR_i = FDR^P - FDR_i^{NP} = \frac{FDR - FDR_i^{CF}}{w} \tag{4}$$

Table 10 presents the actual and counterfactual FDR estimates, and the increase in FDR due to p-hacking using the estimate $\widehat{w} = 72.65\%$. Comparing the FDR and FDR_2^{CF} values, we conclude that p-hacking increased the average FDR among all experiments from 33% to 38% at the 90%/10% discovery threshold. Comparing FDR_2^{CF} before and after adding ΔFDR_2, we conclude that p-hacking increased the average FDR among p-hacked experiments from 33% to 40%.

Table 10: The Impact of p-Hacking on the False Discovery Rate

Discovery Threshold	FDR	FDR_1^{CF}	FDR_2^{CF}	ΔFDR_1	ΔFDR_2
90%/10%	38.07%	35.92%	32.77%	2.96%	7.30%

6 What is the cost of inflating the FDR?

Improper optional stopping increased the average FDR among p-hacked experiments from 33% to 40%. This generates two possible costs for a company. The first is a cost of commission. Facing a false discovery, the company will needlessly switch to a new treatment and incur a switching cost. For many experiments this cost may be low, like changing the background color of a webpage. But for some it may be quite substantial, like building and rolling out the infrastructure to enable a new shipping policy.

The second cost of a false discovery is a cost of omission. Erroneously believing to have found an improvement, the company stops further exploring for better treatments. Consequently the company will delay (or completely forego) finding and rolling out a more effective policy. Our data allow us to quantify this cost of omission. Specifically, we quantify the expected improvement in effectiveness if one more experiment were run.

Let θ be the true effect size of an experiment, i.e., the difference in conversion rates between the baseline and the variation, and let $\widehat{\theta}$ be its observed estimate. Assuming that the company switches away from the baseline only if the observed estimate is positive, the expected improvement in effectiveness if one more experiment were run is:

$$Pr(\widehat{\theta} > 0) \cdot \mathbb{E}[\theta|\widehat{\theta} > 0] + [1 - Pr(\widehat{\theta} > 0)] \cdot 0 \tag{5}$$

The assumptions that the company forgoes only a single experiment and that it would implement any treatment with a positive observed estimate regardless of its statistical significance, make Equation 5 a conservative estimate.

Next, we assume that the true effect size θ is zero with probability π_0 and that $\theta \sim \mathcal{N}(\mu, \sigma^2)$ otherwise. Consistent with the central limit theorem, $\widehat{\theta}|\theta \sim \mathcal{N}(\theta, s^2)$. As shown in Online Appendix B:

$$\mathbb{E}[\theta|\widehat{\theta} > 0] = (1 - \pi_0)\frac{\mu\Phi\left(\frac{\mu}{\sqrt{s^2+\sigma^2}}\right) + \frac{\sigma^2}{\sqrt{s^2+\sigma^2}}\phi\left(\frac{\mu}{\sqrt{s^2+\sigma^2}}\right)}{\frac{\pi_0}{2} + (1 - \pi_0)\Phi\left(\frac{\mu}{\sqrt{s^2+\sigma^2}}\right)} \tag{6}$$

When $\pi_0 = 0$, Equation 6 reduces to the conditional expected value in the standard Heckman selection model (Online Appendix B).

The parameters entering Equation 6 can be estimated conveniently using a latent class model of the observed estimates where $\widehat{\theta} \sim \mathcal{N}(0, s^2)$ with probability π_0 and $\widehat{\theta} \sim \mathcal{N}(\mu, s^2 + \sigma^2)$ with probability $1 - \pi_0$. We estimate this model using the effect sizes observed on the day before termination as they are unaffected by p-hacking. As noted earlier we have 2,053 experiments meeting that condition.

The estimates are reported in Table 11. Even though this analysis uses effect sizes rather than p-values as data, the estimate of π_0 is strikingly close to that obtained from the FDR analysis using p-values (76.7% vs. 72.7%). This similarity alleviates a possible concern that the FDR procedure used in Section 5 over-estimates the true null rate because the original experiments are under-powered and ineffective at detecting moderately small non-null effects. Also, the estimate of the measurement error s^2, which is very small, compared to both the mean and the variance of non-null effects (Table 11), suggests sufficient power. The fact that $\widehat{\mu}(1-\widehat{\pi}_0)$ is close to the observed sample mean of $\widehat{\theta}$ (0.0036 vs. 0.0043) adds further credibility to the estimates.

Entering the values from Table 11 into Equation (6) provides an estimate of $\mathbb{E}[\theta|\widehat{\theta} > 0] = 0.0232$. Since 55.9% of the experiments have $\widehat{\theta} > 0$, Equation 5 implies that the expected cost of omission is an effect size of 0.0130. This corresponds to the 58^{th} percentile of the positive observed effect sizes and the 77^{th} percentile of all observed effect sizes ($\widehat{\theta}$).

Table 11: Parameters of the Distribution of Estimated Effect Sizes

	Paramater Estimate	95% C.I.
π_0	0.7665^{***}	[0.7384, 0.7925]
μ	0.0155^{***}	[0.0057, 0.0253]
s^2	0.0001^{***}	[0.0001, 0.0002]
$s^2 + \sigma^2$	0.0117^{***}	[0.0099, 0.0134]

N=2,053. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

To express this cost of omission in terms of lift, which is a more commonly used outcome metric in A/B testing, we denote the conversion rate of the baseline as η and express the expected increase in lift as:

$$Pr(\widehat{\theta} > 0) \cdot \mathbb{E}\left[lift|\widehat{\theta} > 0\right] = Pr(\widehat{\theta} > 0) \cdot \mathbb{E}\left[\frac{\theta}{\eta}|\widehat{\theta} > 0\right] \approx Pr(\widehat{\theta} > 0) \cdot \frac{\mathbb{E}[\theta|\widehat{\theta} > 0]}{\mathbb{E}[\eta|\widehat{\theta} > 0]} \tag{7}$$

Using the previously calculated value for $Pr(\widehat{\theta} > 0) \cdot \mathbb{E}[\theta|\widehat{\theta} > 0]$, and estimating $\mathbb{E}[\eta|\widehat{\theta} > 0]$ as the mean of the conditional empirical distribution (0.665), we estimate the expected cost of omission in terms of lift at 1.95%. This corresponds to the 58^{th} percentile of the positive observed lifts and the 76^{th} percentile of all observed lifts. Hence the expected opportunity cost of omission following a false discovery is a fairly large forgone gain in lift.

7 Conclusion

We investigate to what extent online A/B experimenters p-hack by stopping their experiments based on the p-value of the treatment effect, and how much this behavior negatively affects the diagnosticity and business value of the experiments. We use data on 2,101 experiments and a regression discontinuity design to detect improper optional stopping. Though our setting is a specific A/B testing platform, similar behaviors to those we document likely exist in other types of business experiments allowing for optional stopping. Examples include in-house email experiments, in-house pricing experiments by retailers, and online advertising experiments.

We find evidence of p-hacking, but only for positive effects and only at the 90% confidence threshold. About 73% of experimenters p-hack when their experiments are within the 89.5%±1.13% confidence window. Notably, these findings come from a platform that declared winners at 95%, but not 90% confidence. Hence, our evidence of p-hacking *cannot* be attributed to the visually salient declaration of the platform.

We also investigate the consequences of p-hacking on the false discovery rate (FDR), and forgone improvements in effectiveness. We find that in roughly 73% – 77% of the experiments the effect is truly null, and that the FDR is 38% at the 90%/10% critical levels, 27% at the 95%/5% critical levels, and 11% at the 99%/1% critical levels. As points of reference, the true null rate in academic psychology is estimated to be about 90% (V. Johnson et al. 2017), and FDRs in medical research are believed to range between 20% and 50% (Benjamini and Hechtlinger 2013). Most importantly, we find that p-hacking increased the average 90%/10% FDR among all experiments from 33% to 38%, and the average FDR among p-hacked experiments from 33% to 40%. Assuming that following such a false discovery experimenters stop searching for more effective treatments, we estimate the expected cost of a false discovery as a forgone gain of 1.95% in lift. This corresponds to the 76^{th} percentile of observed lifts.

Experimenters in our data seem to deviate from profit maximization. If the experiments are run to maximize learning about effect sizes while ignoring short term profits, we should not observe p-hacking that inflates FDRs. In contrast, if experiments are run to maximize short-term profits, we should not observe experiments with larger effect sizes being terminated later, as this prevents the most effective intervention from being rolled out quickly (Azevedo et al. 2018, Feit and Berman 2018).

Our study adds to a small body of previous research documenting that the average effect of online interventions is small. Beyond documenting that effects are typically small (a difference of 0.5% in conversion or a lift of 11.2%), we estimate that 73% of the effects in our sample are truly null. The latter supports earlier suspicions that some of the frustrations with A/B testing may stem from the ineffectiveness of the interventions being tested rather than from inadequacies of the method itself (Fung 2014).

A unique characteristic of our data is its panel structure tracking effect sizes and p-values every day throughout the experiment. This offers three benefits. First, when combined with a regression

discontinuity design, this data structure provides stronger causal identification than using only terminal p-values. Second, the panel structure allows us to estimate a mixture model and quantify the fraction of experimenters engaging in p-hacking. Third, our data allows us to use the statistics on the day *before* the experiment was terminated to compute the counterfactual false discovery rate (FDR) without p-hacking and to compare it to the FDR on the day of termination. The latter, in turn, allows us to quantify the cost in forgone learning for the business.

An important limitation of the data is that experimenters may assess effects on multiple dimensions of which we track only engagement. To the extent that experimenters' stopping behavior is based on the p-value on dimensions other than engagement, our findings likely under-estimate the prevalence and magnitude of p-hacking. This may account in part for the lack of evidence of p-hacking at confidence levels other than 90%, but cannot explain the evidence of p-hacking at 90% as being an artefact.

Our findings add urgency to earlier calls for research on experimenters' intentions and decision rules when running and analyzing experiments (e.g., Greenland 2017, Leek et al. 2017). As many have pointed out, the solution to p-hacking is very unlikely to lie in more and better statistical training. Rather, we need to have a better understanding of how and why people actually behave before we try to develop effective solutions to the problem. For instance, one notable but unexpected finding is that experimenters stop their experiments earlier if they find small positive rather than large positive effects. This might stem from a desire to safeguard small but fortuitous positive results by pulling the plug on the experiment before the effect regresses to the mean. Alternatively, it may stem from the desire to move to the next experiment rather than wasting more time testing a small effect. As a third possibility, the pattern may stem from the concern that large positive effects are too big to be true and the expectation that more credible estimates will follow if the experiment is run longer.

Agency considerations in commercial A/B testing are another facet of p-hacking that has received little attention in previous discussions. A/B testers often act as agents, either as employees or as third-party service providers. Consequently they may prefer to show significant results, especially if positive. Future studies in which the agency relationship of the experimenter is known would be able to provide empirical insights on this issue.

There are at least four strategies to address improper optional stopping, and the resulting

high-FDR problem in A/B testing. The first is to tighten the significance threshold, e.g., from 0.05 to 0.005 (Benjamin et al. 2018). Our finding that experimenters did not adhere to the p-value recommended by the platform suggests that this strategy need not be effective. By making significance harder to achieve, "moving the significance goal posts" may even increase the propensity to p-hack. The second strategy, implemented by Optimizely several months after the end of our data window, is to use proper sequential testing and FDR-control procedures, resulting in corrected p-values (Johari et al. 2017). This strategy is not meant to deter peeking at p-values repeatedly or terminating experiments based on the p-value, but is meant to neutralize the harmful effect of such behavior on the FDR. The third strategy is to use Bayesian hypothesis testing robust to optional stopping (Wagenmakers 2007, Deng et al. 2016). The fourth strategy is to forego null hypothesis testing altogether and to approach the business decision that A/B testing is meant to inform as a decision theoretic problem (Feit and Berman 2018), e.g., using multi-armed bandits (Schwartz et al. 2017). We hope that our findings will provide further impetus to develop strategies accommodating problematic experimenter behaviors, asses the effectiveness of these strategies, and help businesses extract greater value from their experiments.

References

Anscombe, F. J. (1954). Fixed-sample-size analysis of sequential observations. *Biometrics 10*(1), 89–100.

Armitage, P., C. McPherson, and B. Rowe (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General) 132*(2), 235–244.

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

Bartroff, J., T. L. Lai, and M.-C. Shih (2012). *Sequential Experimentation in Clinical Trials: Design and Analysis*. New York: Springer-Verlag.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, et al. (2018). Redefine statistical significance. *Nature Human Behaviour 2*(1), 6–10.

Benjamini, Y. and Y. Hechtlinger (2013). Discussion: An estimate of the science-wise false discovery rate and applications to top medical journals by Jager and Leek. *Biostatistics 15*(1), 13–16.

Berger, J. O. and M. Delampady (1987). Testing precise hypotheses. *Statistical Science 2*(3), 317–335.

Bor, J., E. Moscoe, P. Mutevedzi, M.-L. Newell, and T. Bärnighausen (2014). Regression discontinuity designs in epidemiology: Causal inference without randomized trials. *Epidemiology 25*(5), 729–737.

Brodeur, A., N. Cook, A. Heyes, et al. (2018). Methods matter: P-hacking and causal inference in economics. Technical report, IZA - Institute of Labor Economics.

Bruns, S. B. and J. P. Ioannidis (2016). P-curve and p-hacking in observational research. *PLoS One 11*(2), e0149144.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2017). rdrobust: Software for regression discontinuity designs. *Stata Journal 17*(2), 372–404.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica 82*(6), 2295–2326.

Camerer, C. F., A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour 2*(9), 637–644.

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

Feit, E. M. and R. Berman (2018). Profit-maximizing A/B tests. Working Paper.

Fiske, D. W. and L. V. Jones (1954). Sequential analysis in psychological research. *Psychological Bulletin 51*(3), 264–275.

Fung, K. (2014). Yes, A/B testing is still necessary. https://hbr.org/2014/12/yes-ab-testing-is-still-necessary.

Gelman, A. and G. Imbens (2018). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*. In Press.

Gelman, A. and A. Zelizer (2015). Evidence on the deleterious impact of sustained use of polynomial regression on causal inference. *Research & Politics 2*(1), 1–7.

Goodson, M. (2014). Most winning A/B test results are illusory. Technical report, Qubit.

Greenland, S. (2017). Invited commentary: The need for cognitive science in methodology. *American Journal of Epidemiology 186*(6), 639–645.

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

Hodges, J. and E. Lehmann (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological) 16*(2), 261–268.

Horrace, W. C. and R. L. Oaxaca (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters 90*(3), 321–327.

Imbens, G. and K. Kalyanaraman (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies 79*(3), 933–959.

Imbens, G. W. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics 142*(2), 615–635.

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

Johari, R., P. Koomen, L. Pekelis, and D. Walsh (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*, pp. 1517–1525. ACM.

John, L. K., G. Loewenstein, and D. Prelec (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science 23*(5), 524–532.

Johnson, Garrett A., R. A. Lewis, and E. I. Nubbemeyer (2017). The online display ad effectiveness funnel & carryover: Lessons from 432 field experiments. Working Paper.

Johnson, Valen E., R. D. Payne, T. Wang, A. Asher, and S. Mandal (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association 112*(517), 1–10.

Kareev, Y. and Y. Trope (2011). Correct acceptance weighs more than correct rejection: A decision bias induced by question framing. *Psychonomic bulletin & review 18*(1), 103–109.

Leahey, E. (2005). Alphas and asterisks: The development of statistical significance testing standards in sociology. *Social Forces 84*(1), 1–24.

Leek, J., B. B. McShane, A. Gelman, D. Colquhoun, M. B. Nuijten, and S. N. Goodman (2017). Five ways to fix statistics. *Nature 551*(7682), 557–559.

Lewis, R. A. and J. M. Rao (2015). The unfavorable economics of measuring the returns to advertising. *Quarterly Journal of Economics 130*(4), 1941–1973.

Masson, M. E. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods 43*(3), 679–690.

McShane, B. B. and D. Gal (2015). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science 62*(6), 1707–1718.

McShane, B. B. and D. Gal (2017). Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association 112*(519), 885–895.

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.

Pekelis, L., D. Walsh, and R. Johari (2015). The new Stats Engine. Technical report, Optimizely.

Rouder, J. N. (2014). Optional stopping: No problem for bayesians. *Psychonomic Bulletin & Review 21*(2), 301–308.

Schwartz, E. M., E. T. Bradlow, and P. S. Fader (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science 36*(4), 500–522.

Scott, S. L. (2015). Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry 31*(1), 37–45.

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

Simonsohn, U., L. D. Nelson, and J. P. Simmons (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General 143*(2), 534–547.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(3), 479–498.

Storey, J. D. and R. Tibshirani (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences 100*(16), 9440–9445.

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

Van den Bulte, C. and R. Iyengar (2011). Tricked by truncation: Spurious duration dependence and social contagion in hazard models. *Marketing Science 30*(2), 233–248.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review 14*(5), 779–804.

Wald, A. (1947). *Sequential Analysis*. New York: Wiley.

Yu, E. C., A. M. Sprenger, R. P. Thomas, and M. R. Dougherty (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review 21*(2), 268–282.

Online Appendix

A Supplemental Analyses

Table A1: Hazard Regression Results for Different Codings of Lift

	(1) $+1/-1$	(2) Actual Centered	(3) Actual Normalized	(4) $+1/-1$ + Controls	(5) Actual Centered + Controls	(6) Actual Normalized + Controls
Confidence						
0–0.1	-3.3453^{***}	-3.4014^{***}	-3.4014^{***}	-4.3611^{***}	-4.1970^{***}	-4.1970^{***}
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
0.1–0.2	-3.4056^{***}	-3.5057^{***}	-3.5057^{***}	-4.3498^{***}	-4.2338^{***}	-4.2338^{***}
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
0.2–0.3	-3.3292^{***}	-3.4324^{***}	-3.4324^{***}	-4.2563^{***}	-4.1437^{***}	-4.1437^{***}
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
0.3–0.4	-3.4927^{***}	-3.6199^{***}	-3.6199^{***}	-4.4421^{***}	-4.3541^{***}	-4.3541^{***}
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
0.4–0.5	-3.4212^{***}	-3.4867^{***}	-3.4867^{***}	-4.3922^{***}	-4.2364^{***}	-4.2364^{***}
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
0.5–0.6	-3.5222^{***}	-3.6726^{***}	-3.6726^{***}	-4.4606^{***}	-4.3916^{***}	-4.3916^{***}
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
0.6–0.7	-3.4880^{***}	-3.6597^{***}	-3.6597^{***}	-4.4327^{***}	-4.3815^{***}	-4.3815^{***}
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
0.7–0.8	-3.3768^{***}	-3.5780^{***}	-3.5780^{***}	-4.3384^{***}	-4.3204^{***}	-4.3204^{***}
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
0.8–0.9	-3.3572^{***}	-3.6506^{***}	-3.6506^{***}	-4.3303^{***}	-4.3947^{***}	-4.3947^{***}
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
0.9–1	-3.3230^{***}	-3.3365^{***}	-3.3365^{***}	-4.3195^{***}	-4.0964^{***}	-4.0964^{***}
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
0–0.1 × lift	0.1168^{*}	2.7366^{***}	0.4454^{***}	0.0304	2.1386^{***}	0.3481^{***}
	(0.0638)	(0.0000)	(0.0000)	(0.6630)	(0.0016)	(0.0016)
0.1–0.2 × lift	0.2562^{***}	9.1486^{***}	0.6601^{***}	0.1730^{*}	7.7388^{***}	0.5583^{***}
	(0.0041)	(0.0001)	(0.0001)	(0.0658)	(0.0013)	(0.0013)
0.2–0.3 × lift	0.1928^{**}	13.1852^{***}	0.6855^{***}	0.1035	10.6047^{***}	0.5514^{***}
	(0.0274)	(0.0001)	(0.0001)	(0.2586)	(0.0017)	(0.0017)
0.3–0.4 × lift	0.1966^{**}	22.5850^{***}	0.7297^{***}	0.1074	19.2008^{***}	0.6203^{***}
	(0.0359)	(0.0002)	(0.0002)	(0.2733)	(0.0020)	(0.0020)
0.4–0.5 × lift	0.2285^{***}	37.6489^{***}	0.5397^{***}	0.1845^{**}	31.8015^{**}	0.4559^{**}
	(0.0095)	(0.0031)	(0.0031)	(0.0449)	(0.0126)	(0.0126)
0.5–0.6 × lift	-0.1584^{*}	-66.7175^{***}	-0.8723^{***}	-0.1113	-62.8951^{***}	-0.8223^{***}
	(0.0645)	(0.0002)	(0.0002)	(0.2128)	(0.0007)	(0.0007)
0.6–0.7 × lift	-0.2322^{***}	-25.1922^{***}	-1.0487^{***}	-0.1652^{*}	-22.2407^{***}	-0.9258^{***}
	(0.0081)	(0.0000)	(0.0000)	(0.0720)	(0.0003)	(0.0003)
0.7–0.8 × lift	-0.2860^{***}	-16.0022^{***}	-1.0622^{***}	-0.2137^{**}	-14.5900^{***}	-0.9685^{***}

	(0.0004)	(0.0000)	(0.0000)	(0.0125)	(0.0000)	(0.0000)
0.8–0.9 × lift	−0.2148***	−9.1140***	−3.3826***	−0.1559**	−8.2438***	−3.0596***
	(0.0043)	(0.0000)	(0.0000)	(0.0497)	(0.0000)	(0.0000)
0.9–1 × lift	−0.1924***	−0.0206	−0.1770	−0.1304**	−0.0056	−0.0477
	(0.0004)	(0.1927)	(0.1927)	(0.0298)	(0.7273)	(0.7273)
Sample Size				−0.0000	−0.0000	−0.0000
				(0.2141)	(0.2731)	(0.2731)
Past # Experiments				−0.0001	−0.0002	−0.0002
				(0.2121)	(0.1729)	(0.1729)
Industry						
Financial Services				0.0541	0.0220	0.0220
				(0.8595)	(0.9412)	(0.9412)
Gaming				−2.5114*	−2.4928*	−2.4928*
				(0.0555)	(0.0504)	(0.0504)
High Tech				0.0755	0.0834	0.0834
				(0.6440)	(0.6013)	(0.6013)
Insurance				−0.5015	−0.5500	−0.5500
				(0.4155)	(0.3622)	(0.3622)
Media				0.5616***	0.5282***	0.5282***
				(0.0014)	(0.0020)	(0.0020)
Mobile Only				−0.4200	−0.5082	−0.5082
				(0.7922)	(0.7448)	(0.7448)
Non-Profit				0.7611*	0.6847	0.6847
				(0.0926)	(0.1203)	(0.1203)
Other				0.0000	0.0000	0.0000
				(.)	(.)	(.)
Professional Services				−0.3707*	−0.3071*	−0.3071*
				(0.0512)	(0.0985)	(0.0985)
Retail				0.0941	0.0684	0.0684
				(0.5382)	(0.6459)	(0.6459)
Schools & Education				−0.1492	−0.1441	−0.1441
				(0.6036)	(0.6082)	(0.6082)
Telecommunications				0.9143**	0.8583**	0.8583**
				(0.0296)	(0.0366)	(0.0366)
Travel & Entertainment				0.1159	0.0920	0.0920
				(0.5914)	(0.6622)	(0.6622)
Day of the Week						
Sunday				−2.2420***	−2.2414***	−2.2414***
				(0.0000)	(0.0000)	(0.0000)
Monday				0.0000	0.0000	0.0000
				(.)	(.)	(.)
Tuesday				−0.0875	−0.0927	−0.0927
				(0.2230)	(0.1967)	(0.1967)
Wednesday				−0.1755**	−0.1779**	−0.1779**
				(0.0184)	(0.0169)	(0.0169)
Thursday				−0.0254	−0.0303	−0.0303
				(0.7282)	(0.6787)	(0.6787)

Friday	-0.3014^{***}	-0.3059^{***}	-0.3059^{***}
	(0.0001)	(0.0001)	(0.0001)
Saturday	-2.1629^{***}	-2.1700^{***}	-2.1700^{***}
	(0.0000)	(0.0000)	(0.0000)
Day in Experiment			
1	0.0000	0.0000	0.0000
	(.)	(.)	(.)
2	0.6136^{***}	0.5079^{**}	0.5079^{**}
	(0.0020)	(0.0109)	(0.0109)
3	0.9246^{***}	0.8019^{***}	0.8019^{***}
	(0.0000)	(0.0001)	(0.0001)
4	0.4707^{**}	0.3357	0.3357
	(0.0374)	(0.1380)	(0.1380)
5	0.7077^{***}	0.5561^{**}	0.5561^{**}
	(0.0012)	(0.0110)	(0.0110)
6	1.0084^{***}	0.8491^{***}	0.8491^{***}
	(0.0000)	(0.0000)	(0.0000)
7	1.1306^{***}	0.9590^{***}	0.9590^{***}
	(0.0000)	(0.0000)	(0.0000)
8	1.3717^{***}	1.1909^{***}	1.1909^{***}
	(0.0000)	(0.0000)	(0.0000)
9	1.2931^{***}	1.1106^{***}	1.1106^{***}
	(0.0000)	(0.0000)	(0.0000)
10	1.3157^{***}	1.1279^{***}	1.1279^{***}
	(0.0000)	(0.0000)	(0.0000)
11	1.0637^{***}	0.8688^{***}	0.8688^{***}
	(0.0000)	(0.0002)	(0.0002)
12	1.4004^{***}	1.1998^{***}	1.1998^{***}
	(0.0000)	(0.0000)	(0.0000)
13	1.2878^{***}	1.0860^{***}	1.0860^{***}
	(0.0000)	(0.0000)	(0.0000)
14	1.6624^{***}	1.4685^{***}	1.4685^{***}
	(0.0000)	(0.0000)	(0.0000)
15	1.6283^{***}	1.4265^{***}	1.4265^{***}
	(0.0000)	(0.0000)	(0.0000)
16	1.1583^{***}	0.9614^{***}	0.9614^{***}
	(0.0000)	(0.0001)	(0.0001)
17	1.2919^{***}	1.0937^{***}	1.0937^{***}
	(0.0000)	(0.0000)	(0.0000)
18	1.3933^{***}	1.1772^{***}	1.1772^{***}
	(0.0000)	(0.0000)	(0.0000)
19	1.5351^{***}	1.3162^{***}	1.3162^{***}
	(0.0000)	(0.0000)	(0.0000)
20	1.7254^{***}	1.5099^{***}	1.5099^{***}
	(0.0000)	(0.0000)	(0.0000)
21	1.5189^{***}	1.2934^{***}	1.2934^{***}
	(0.0000)	(0.0000)	(0.0000)
22	1.7661^{***}	1.5324^{***}	1.5324^{***}

	(0.0000)	(0.0000)	(0.0000)
23	0.9285***	0.6927**	0.6927**
	(0.0020)	(0.0210)	(0.0210)
24	1.0638***	0.8330***	0.8330***
	(0.0004)	(0.0057)	(0.0057)
25	2.1345***	1.9100***	1.9100***
	(0.0000)	(0.0000)	(0.0000)
26	1.9545***	1.7190***	1.7190***
	(0.0000)	(0.0000)	(0.0000)
27	1.5070***	1.2753***	1.2753***
	(0.0000)	(0.0000)	(0.0000)
28	1.7804***	1.5480***	1.5480***
	(0.0000)	(0.0000)	(0.0000)
29	1.0640***	0.8276***	0.8276***
	(0.0007)	(0.0085)	(0.0085)
30	1.2986***	1.0671***	1.0671***
	(0.0000)	(0.0004)	(0.0004)
31	1.4782***	1.2501***	1.2501***
	(0.0000)	(0.0000)	(0.0000)
32	1.8248***	1.5948***	1.5948***
	(0.0000)	(0.0000)	(0.0000)
33	0.9272**	0.6887*	0.6887*
	(0.0205)	(0.0851)	(0.0851)
34	1.7255***	1.4774***	1.4774***
	(0.0000)	(0.0000)	(0.0000)
35	1.6361***	1.3912***	1.3912***
	(0.0000)	(0.0000)	(0.0000)
36–40	1.6210***	1.3717***	1.3717***
	(0.0000)	(0.0000)	(0.0000)
41–45	1.5336***	1.2805***	1.2805***
	(0.0000)	(0.0000)	(0.0000)
46–50	1.5804***	1.3239***	1.3239***
	(0.0000)	(0.0000)	(0.0000)
51–55	1.8873***	1.6214***	1.6214***
	(0.0000)	(0.0000)	(0.0000)
56–60	1.6472***	1.3755***	1.3755***
	(0.0000)	(0.0000)	(0.0000)
61–65	1.6289***	1.3557***	1.3557***
	(0.0000)	(0.0000)	(0.0000)
66–70	1.6408***	1.3684***	1.3684***
	(0.0000)	(0.0000)	(0.0000)
71–75	2.2798***	1.9953***	1.9953***
	(0.0000)	(0.0000)	(0.0000)
76–80	1.2333***	0.9592***	0.9592***
	(0.0003)	(0.0048)	(0.0048)
81–85	2.3425***	2.0512***	2.0512***
	(0.0000)	(0.0000)	(0.0000)
86–90	1.0673***	0.7806*	0.7806*
	(0.0088)	(0.0553)	(0.0553)

91–95				2.2794***	1.9934***	1.9934***
				(0.0000)	(0.0000)	(0.0000)
96–100				1.9053***	1.6047***	1.6047***
				(0.0000)	(0.0000)	(0.0000)
101–150				2.3760***	2.0508***	2.0508***
				(0.0000)	(0.0000)	(0.0000)
151–200				3.0743***	2.7238***	2.7238***
				(0.0000)	(0.0000)	(0.0000)
201–250				2.5230***	2.1538***	2.1538***
				(0.0000)	(0.0000)	(0.0000)
251+				2.6809***	2.3316***	2.3316***
				(0.0000)	(0.0000)	(0.0000)
LL	−9257.273	−9192.734	−9192.734	−8696.030	−8653.377	−8653.377
σ	0.752	0.765	0.765	1.151	1.110	1.110

Experiment-Days = 76,215; # Experiments = 2,101; # Experimenters = 916; p-values in parentheses.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Figure A1: Histograms of average values of covariates inside 0.001 confidence-wide bins around different critical values of confidence. Left column: critical value=.895, Right column: critical value=.945. Covariates (from top to bottom): Day in the experiment, Lift, Number of Visitors, Weekend Indicator

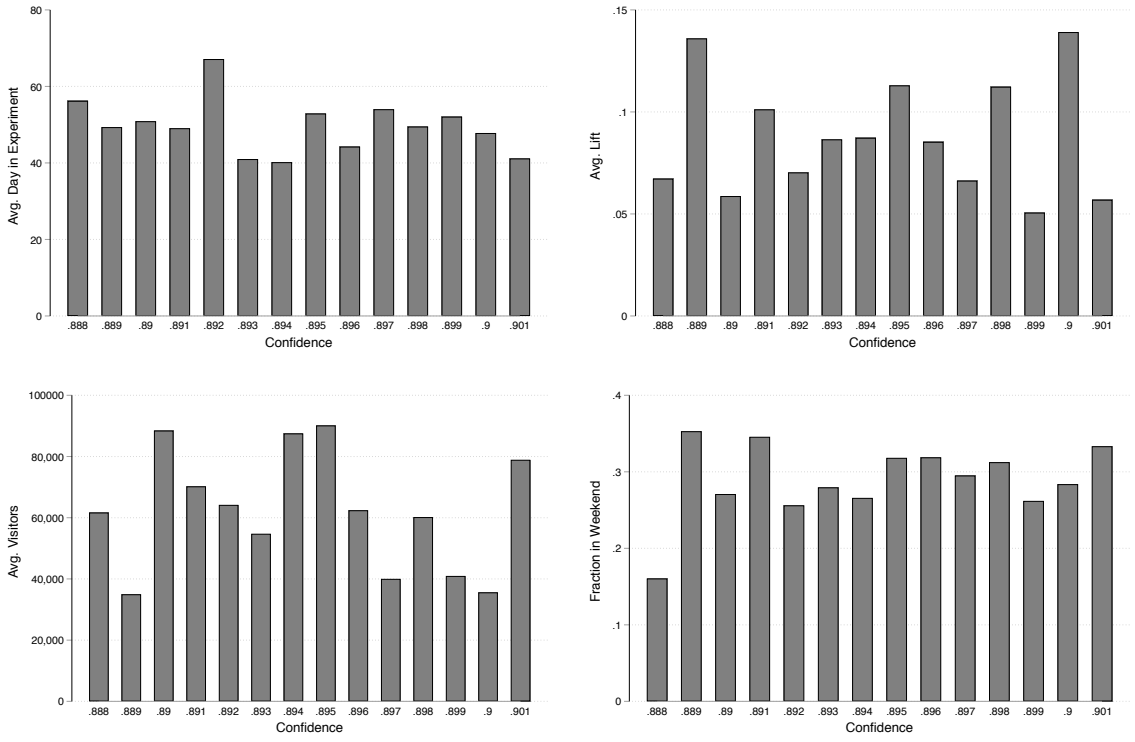


Table A2: Regression discontinuity tests for covariates

	Day in Experiment		
	Effect	p-value	95% C.I.
Linear	12.2191	0.3199	[−11.8598, 36.2981]
Linear, Bias Corr.	14.5629	0.2587	[−10.7061, 39.8319]
	Lift		
	Effect	p-value	95% C.I.
Linear	0.0165	0.5503	[−0.0377, 0.0708]
Linear, Bias Corr.	0.0163	0.5879	[−0.0426, 0.0751]
	Total Visitors		
	Effect	p-value	95% C.I.
Linear	15777.23	0.8003	[−106478.62, 138033.08]
Linear, Bias Corr.	7245.71	0.9171	[−129211.59, 143703.00]
	Weekend Indicator		
	Effect	p-value	95% C.I.
Linear	0.0561	0.3967	[−0.0736, 0.1857]
Linear, Bias Corr.	0.0675	0.3671	[−0.0792, 0.2142]

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A3: Latent class model estimates

	Model (1)		Model (2)	
Response equations	non p-hacking	p-hacking	non p-hacking	p-hacking
Intercept	-67238.24^{***}	-3.786^{***}	-64385.13^{***}	-3.779^{***}
	(617.00)	(0.707)	(583.36)	(0.709)
X	-265396.15^{***}	-4.203	-254135.12^{***}	-4.250
	(1005.49)	(2.707)	(2529.98)	(2.715)
X^2	-261698.13^{***}	-3.968^{*}	-250595.18^{***}	-4.015^{*}
	(809.66)	(2.341)	(5308.35)	(2.351)
D		1.848^{**}		1.840^{**}
		(0.848)		(0.850)
$D \cdot X$		-1.333		1.246
		(3.504)		(3.510)
$D \cdot X^2$		8.669^{***}		8.698^{***}
		(3.022)		(3.035)
Class membership equation				
Intercept	-0.977^{**}		-1.006^{*}	
	(0.433)		(0.573)	
High-Tech			1.020	
			(0.813)	
Media			-0.872	
			(1.271)	
Retail			-0.535	
			(2.028)	
log(Past # Experiments)			-0.013	
			(0.228)	
LL	-257.331		-255.776	

Standard Errors in Parentheses. $N = 1894$. * $p < 0.1$ ** $p < 0.05$ *** $p < 0.01$.
To facilitate presentation, confidence levels were converted to percentage points.
Past # of experiments is log-transformed to achieve model convergence.
Note: the class membership equation models the probability of not being a p-hacker.

B Derivations for Section 6

B.1 Derivation of Equation 6

Assume that:

$$\theta = 0 \text{ with probability } \pi_0$$
$$\theta \sim N(\mu, \sigma^2) \text{ with probability } 1 - \pi_0$$
$$\widehat{\theta} = \theta + \varepsilon \text{ where } \varepsilon \sim N(0, s^2)$$

The PDF of $\widehat{\theta}$ conditional on θ is:

$$f(\widehat{\theta}|\theta) = \frac{1}{s}\phi\left(\frac{\widehat{\theta} - \theta}{s}\right) \tag{B1}$$

and the unconditional PDF is:

$$f(\widehat{\theta}) = \int_\theta f(\widehat{\theta}|\theta) Pr(\theta) d\theta = \pi_0 \frac{1}{s}\phi\left(\frac{\widehat{\theta}}{s}\right) + (1 - \pi_0)\frac{1}{\sqrt{s^2 + \sigma^2}}\phi\left(\frac{\widehat{\theta} - \mu}{\sqrt{s^2 + \sigma^2}}\right) \tag{B2}$$

We now derive $\mathbb{E}[\theta|\widehat{\theta} > 0]$:

$$\mathbb{E}[\theta|\widehat{\theta} > 0] = \frac{\int_{\widehat{\theta}>0} \mathbb{E}[\theta|\widehat{\theta}] f(\widehat{\theta}) d\widehat{\theta}}{\int_{\widehat{\theta}>0} f(\widehat{\theta}) d\widehat{\theta}} \tag{B3}$$
$$= \frac{\int_{\widehat{\theta}>0} \int_\theta \theta f(\theta|\widehat{\theta}) d\theta f(\widehat{\theta}) d\widehat{\theta}}{\int_{\widehat{\theta}>0} f(\widehat{\theta}) d\widehat{\theta}} \tag{B4}$$
$$= \int_\theta \theta f(\theta) \frac{\int_{\widehat{\theta}>0} f(\widehat{\theta}|\theta) d\widehat{\theta}}{\int_{\widehat{\theta}>0} f(\widehat{\theta}) d\widehat{\theta}} d\theta \tag{B5}$$
$$= \int_\theta \theta f(\theta) \frac{\Phi\left(\frac{\theta}{s}\right)}{\int_{\widehat{\theta}>0} f(\widehat{\theta}) d\widehat{\theta}} d\theta \tag{B6}$$
$$= (1 - \pi_0) \frac{\int_\theta \theta \frac{1}{\sigma}\phi\left(\frac{\theta - \mu}{\sigma}\right) \Phi\left(\frac{\theta}{s}\right) d\theta}{\int_{\widehat{\theta}>0} f(\widehat{\theta}) d\widehat{\theta}} \tag{B7}$$

Letting $y = \frac{\theta-\mu}{\sigma}$:

$$\int_\theta \theta \frac{1}{\sigma}\phi\left(\frac{\theta-\mu}{\sigma}\right)\Phi\left(\frac{\theta}{s}\right)d\theta = \int_y (\sigma y+\mu)\phi(y)\Phi\left(\frac{y+\frac{\mu}{\sigma}}{\frac{s}{\sigma}}\right)dy \tag{B8}$$

and using:

$$\int_y y\phi(y)\Phi\left(\frac{y+b}{a}\right)dy = \frac{1}{\sqrt{a^2+1}}\phi\left(\frac{b}{\sqrt{a^2+1}}\right) \tag{B9}$$

$$\int_y \phi(y)\Phi\left(\frac{y+b}{a}\right)dy = \Phi\left(\frac{b}{\sqrt{a^2+1}}\right) \tag{B10}$$

we have:

$$\int_\theta \theta \frac{1}{\sigma}\phi\left(\frac{\theta-\mu}{\sigma}\right)\Phi\left(\frac{\theta}{s}\right)d\theta = \frac{\sigma^2}{\sqrt{s^2+\sigma^2}}\phi\left(\frac{\mu}{\sqrt{s^2+\sigma^2}}\right)+\mu\Phi\left(\frac{\mu}{\sqrt{s^2+\sigma^2}}\right) \tag{B11}$$

and hence we can rewrite Equation B7 as:

$$\mathbb{E}[\theta|\widehat{\theta}>0] = (1-\pi_0)\frac{\frac{\sigma^2}{\sqrt{s^2+\sigma^2}}\phi\left(\frac{\mu}{\sqrt{s^2+\sigma^2}}\right)+\mu\Phi\left(\frac{\mu}{\sqrt{s^2+\sigma^2}}\right)}{\int_{\widehat{\theta}>0} f(\widehat{\theta})d\widehat{\theta}} \tag{B12}$$

Finally:

$$\int_{\widehat{\theta}>0} f(\widehat{\theta})d\widehat{\theta} = \frac{\pi_0}{2}+(1-\pi_0)\Phi\left(\frac{\mu}{\sqrt{s^2+\sigma^2}}\right) \tag{B13}$$

Combining Equations B12 and B13 results in Equation 6:

$$\mathbb{E}[\theta|\widehat{\theta}>0] = (1-\pi_0)\frac{\mu\Phi\left(\frac{\mu}{\sqrt{s^2+\sigma^2}}\right)+\frac{\sigma^2}{\sqrt{s^2+\sigma^2}}\phi\left(\frac{\mu}{\sqrt{s^2+\sigma^2}}\right)}{\frac{\pi_0}{2}+(1-\pi_0)\Phi\left(\frac{\mu}{\sqrt{s^2+\sigma^2}}\right)} \tag{B14}$$

B.2 Similarity of Equation 6 to the Heckman correction

When $\pi_0 = 0$, our model can be expressed as:

$$\theta = \mu + u_1 \tag{B15}$$

$$\widehat{\theta} = \mu + u_2 > 0 \tag{B16}$$

where $u_1 \sim \mathcal{N}(0, \sigma^2)$, $u_2 \sim \mathcal{N}(0, s^2 + \sigma^2)$, $\rho = corr(u_1, u_2) = \frac{\sigma}{\sqrt{s^2+\sigma^2}}$.
If one considers θ only if $\widehat{\theta} > 0$, then the selection corrected expected value of θ equals:

$$\mathbb{E}[\theta|\widehat{\theta} > 0] = \mu + \sigma \cdot \rho \cdot \frac{\phi\left(\frac{\mu}{\sqrt{s^2+\sigma^2}}\right)}{\Phi\left(\frac{\mu}{\sqrt{s^2+\sigma^2}}\right)} \tag{B17}$$

Substituting $\rho = \frac{\sigma}{\sqrt{s^2+\sigma^2}}$ into this standard Heckman correction, we obtain Equation 6 with $\pi_0 = 0$:

$$\mathbb{E}[\theta|\widehat{\theta} > 0] = \mu + \frac{\sigma^2}{\sqrt{s^2+\sigma^2}} \frac{\phi\left(\frac{\mu}{\sqrt{s^2+\sigma^2}}\right)}{\Phi\left(\frac{\mu}{\sqrt{s^2+\sigma^2}}\right)} \tag{B18}$$