Marketing Science Institute Working Paper Series 2018
Report No. 18-130

p-Hacking and False Discovery in A/B Testing

Ron Berman, Leonid Pekelis, Aisling Scott, and Christophe Van den Bulte

Report Summary

Ron Berman, Leonid Pekelis, Aisling Scott, and Christophe Van den Bulte investigate whether online A/B experimenters "p-hack" by stopping their experiments based on the p-value of the treatment effect. If A/B testers p-hack, this behavior may inflate the number of experiments that show significance but actually have no effect (are null), resulting in incorrect decisions and foregone earnings for the company.

The Study

Their data contains 2,101 commercial experiments from the Optimizely platform, before the platform implemented a new method that alleviates the negative effects of p-hacking. In the data, experimenters can track the magnitude and significance level of the effect every day of the experiment. They use a regression discontinuity design to detect p-hacking, i.e., the causal effect of reaching a particular p-value on stopping behavior.

They find that experimenters indeed p-hack, especially for positive effects. Specifically, about 57% of experimenters p-hack when the experiment reaches 90% confidence, and p-hacking increases the false discovery rate (FDR) from 33% to 42% among experiments p-hacked at 90% confidence. Furthermore, approximately 70% of the effects of all experiments are truly null, i.e., not expected to yield any improvement over the baseline.

Assuming that false discoveries cause experimenters to stop exploring for more effective treatments, the authors estimate the expected cost of a false discovery to be a loss of 1.95% in lift (compared to a median lift of 11% in all experiments), which corresponds to the 76th percentile of observed lifts.

Put into Practice

These findings have implications for practitioners and platform designers.

- For A/B testers, the authors estimate the costs and identify the source of potential pitfalls that companies should avoid when designing and running experiments. For example, the finding that 70% of treatment effects are null should emphasize to experimenters how difficult it is to design new and effective online innovations, and that they should not be fooled if they find significant effects too often.

- Furthermore, because agency considerations may be at play, experimenters who use a third party company to design and run experiments for them should be vigilant when receiving reports about the outcomes of these experiments.

- For platform designers, they uncover how the behavior of experimenters affects the efficacy of their platform, and consequently their customer satisfaction.

The authors discuss several potential causes for this behavior and the potential of different methods to remedy these issues for experimenters.

Ron Berman is an Assistant Professor of Marketing, The Wharton School, Leonid Pekelis is a Statistician, OpenDoor, Aisling Scott is a Research Scientist, Facebook, and Christophe Van den Bulte is the Gayfryd Steinberg Professor and Professor of Marketing, The Wharton School.

Acknowledgments
The authors thank Optimizely for making the data available for this project. They also thank Eric Bradlow, Elea McDonnell Feit, Raghuram Iyengar, Pete Koomen, and participants at the 2018 INFORMS Marketing Science Conference for feedback.

1 Introduction

Marketers increasingly turn to online experiments to inform their decisions. This shift is facilitated by various testing platforms like 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 provide standard tools for null hypothesis statistical testing. However, industry observers have pointed to the need to use these tools properly. Specifically, if one repeatedly checks ongoing experiments for significance, results that are declared to be statistically significant may very well not be (Miller 2010, Goodson 2014, Johari et al. 2017).

In response to these concerns, 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 sequential testing (e.g., Pekelis et al. 2015) and multi-armed bandit designs (e.g., Scott 2015 and 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), where difficulties in reproducing research findings have been attributed to two main sources. The first is the bias among many journal editors and research sponsors in favor of novel and statistically significant results. The second is the tendency of authors to engage in behaviors generating significant effects where none exist. Examples of these behaviors, generally referred to as p-hacking, include continuously monitoring the experiment and stopping it once the observed p-value falls below a threshold (usually 0.05), testing many hypotheses but reporting only those that fall below the significance threshold, and excluding participants or transforming the data to get a p-value below the threshold (Simmons et al. 2011, Head et al. 2015).

Academics may p-hack for two reasons. First, facing editorial bias favoring statistical significance, researchers may engage in willful attempts to report significant effects where none exist. Second, experimenters may lack statistical skill. 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. Similarly, researchers who attempt to avoid wasting time or subjects may erroneously believe it is appropriate to test repeatedly for significance and stop the experiment as soon as it achieves significance (Miller 2010, John et al. 2012, Johari et al. 2017).

We investigate whether p-hacking, documented previously in academic research, also exists in commercial A/B testing, and how it harms the diagnosticity of these tests. Specifically, we investigate (1) whether online A/B experimenters p-hack by stopping their experiments based on the p-value of the treatment effect, and (2) how such p-hacking affects the tendency for A/B tests to produce falsely significant results. Our focus is solely on p-hacking in the form of stopping experiments based on the obtained p-value, and how this affects the false discovery rate (FDR).

Why would marketers and other business people engage in p-hacking? Unlike academics, they do not operate under "publish or perish" and hence 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.

We believe that the two main reasons for p-hacking in academic research may also apply to business experiments. First, even when malicious intent is not at play, experimenters may p-hack simply because of poor statistical skills. Many experimenters do not have the background or experience to validly interpret the statistical results provided by a platform. Moreover, A/B testing platforms often make recommendations that do not control for checking the p-value repeatedly during the experiment or for multiple hypothesis testing. The second reason why business experimenters may p-hack is that they often act as agents with incentives to produce significant results. For example, when an outside advertising agency is being contracted to assess the effectiveness of campaigns they designed, they will have strong incentives to report significant positive results if that helps generate more business. Similarly, employees in charge of running A/B experiments and concerned about their perceived competence may benefit in the short term from reporting significant results.

Our study makes five contributions. (1) We document the existence and quantify the prevalence of p-hacking 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 p-hacking 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.

We detect p-hacking by applying a regression discontinuity design (RDD) to panel data of 2,101 experiments run on the Optimizely A/B testing platform in 2014. We estimate that 57% of the experimenters with observations around the 90% confidence level engage in p-hacking. The data is unique in that it has daily observations about the performance and termination of each experiment, rather than only the 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, our data allows us to directly infer the relationship between achieved p-values and stopping behavior, including discontinuities within very narrow intervals around critical p-values. In essence, our data allows us to "look over the shoulder" of the experimenters and draw stronger conclusions about their behavior.

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%. 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 lower than the 90% true null rate documented in academic psychological research (V. Johnson et al. 2017). Our estimate that 73% 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 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 42%. In other words, p-hacking boosts the probability that an effect declared significant is actually a null effect from 33% to 42%, and by doing so greatly 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 42%, we also quantify the expected cost of a false discovery. Assuming that experimenters stop exploring

for more effective treatments after having made 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.

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 in 2014. 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 log-in to the platform's dashboard and view statistics about the experiment at any time during and after the experiment.[1] The experimenter may terminate the experiment 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.

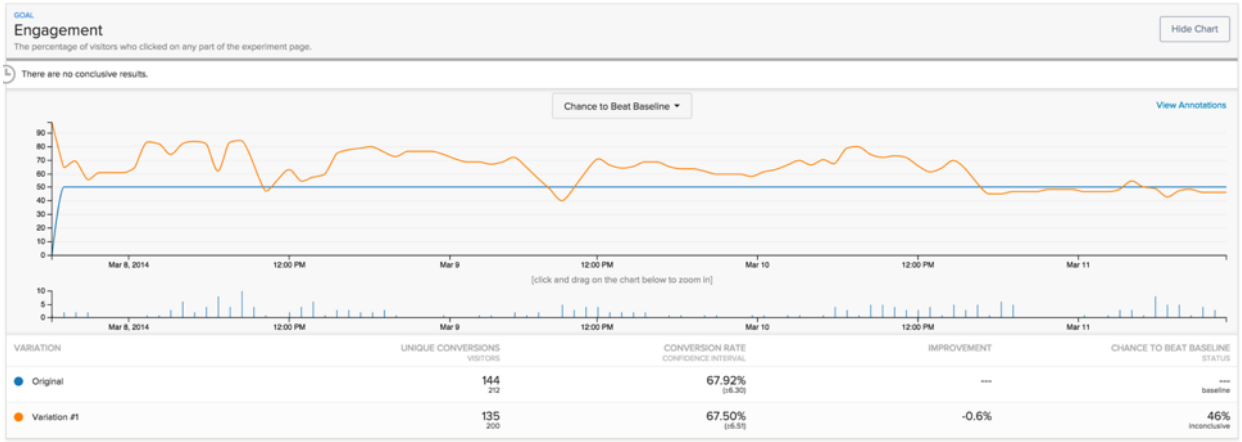
3. Improvement (in percent) – the relative percent difference in conversion rates between the variation and the baseline. This is also known as the Lift of the variation.

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 displays the probability that a specific variation beats the baseline. The metric 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:

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



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

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.

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.

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.

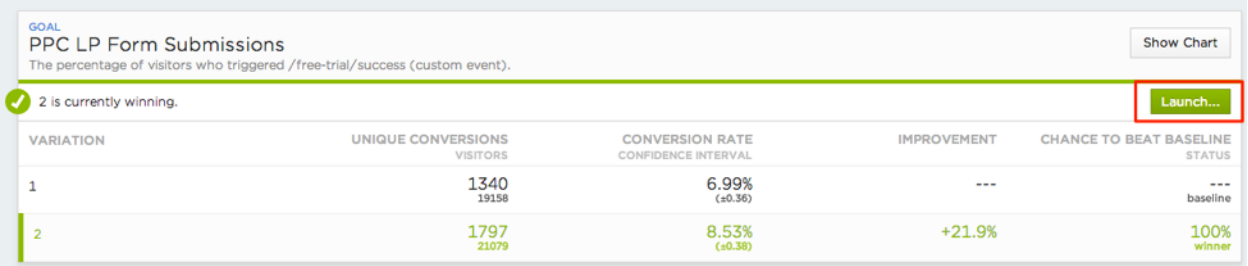
2.3 Set of experiments studied

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 the 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

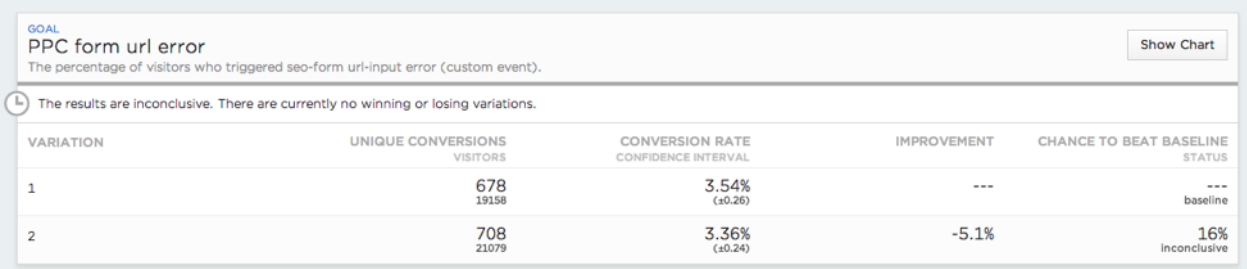
Figure 2: Experimenter Dashboard: Winning, Losing and Inconclusive Variations

(a) Winning Variation



(b) Losing Variation

(c) Inconclusive Variation



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 webpage appears in all variations. Such pre-tests are recommended by platforms. Experiments with 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. It is possible that experimenters pursued a primary goal other than Engagement. We do not have that information.

Finally, in experiments that had more than one non-baseline variation, 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.

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. On average, experiments had between 2 and 3 variations and between 4 and 5 goals. Experimenters varied quite a bit in the number of prior experiments they had run on the platform, with the average across experiments being 364. On their last day, experiments on average included more than 140,000 visitors and had run for 36 days.

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 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

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: 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.

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.

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

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

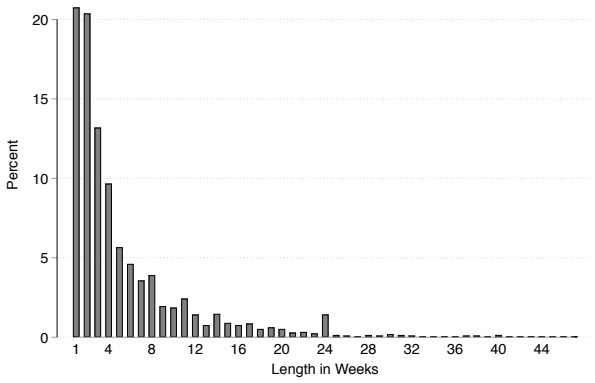
	Start	End
Monday	11.14	21.89
Tuesday	20.32	19.70
Wednesday	22.13	17.85
Thursday	17.71	20.18
Friday	14.85	15.42
Saturday	7.33	2.52
Sunday	6.52	2.43
Total	100.00	100.00

$N = 2,101$.

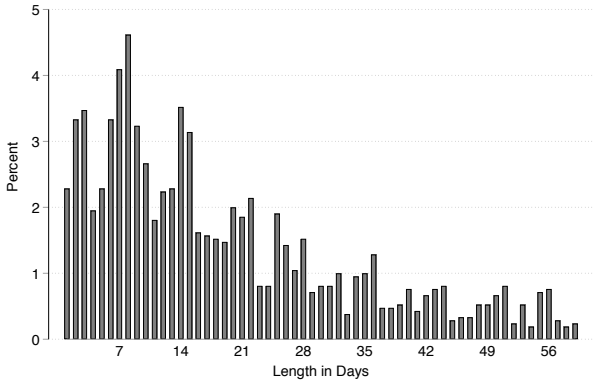
less, and only 5% last longer than 132 days. A small spike in the number of endings between 161 and 168 days is notable, reflecting a tendency not to run experiments beyond 24 weeks. Figure 3b shows the histogram of durations up to 60 days, which covers 82.7% of all experiments. Clearly, there is a downward 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.[2]

Figure 3: Histograms of Experiments' Length

(a) All days



(b) First 60 days



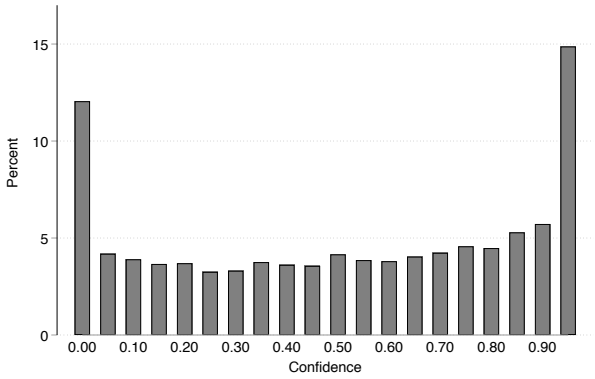
Ending an experiment is also associated with the level of confidence achieved. Figure 4 presents histograms of confidence values on any day during the experiment (4a) and on the day they end (4b). There is slightly more mass at confidence levels of values .85 and higher in Figure 4b than Figure

[2]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.

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).

Figure 4: Histograms of Experiments' Confidence

(a) All days



(b) Last days of experiment

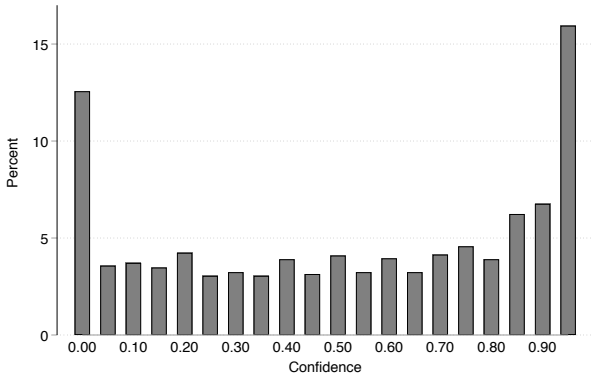


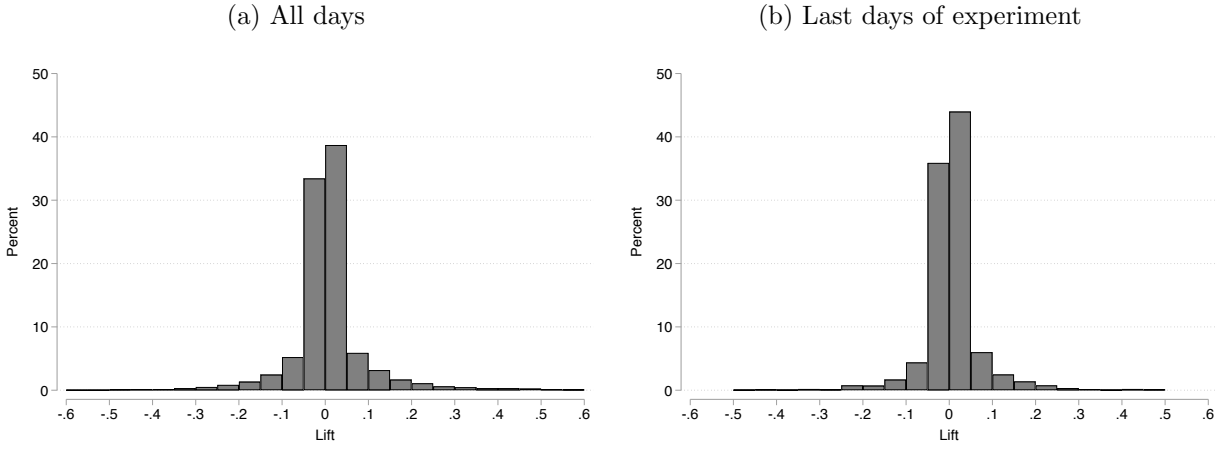
Figure 6 plots the empirical hazard of stopping the experiment on a particular day against the confidence level achieved on that day, binned in 10% increments. The slope of the grey line indicates that when the confidence level is below 0.5 and hence the lift is negative, there is a very slight negative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

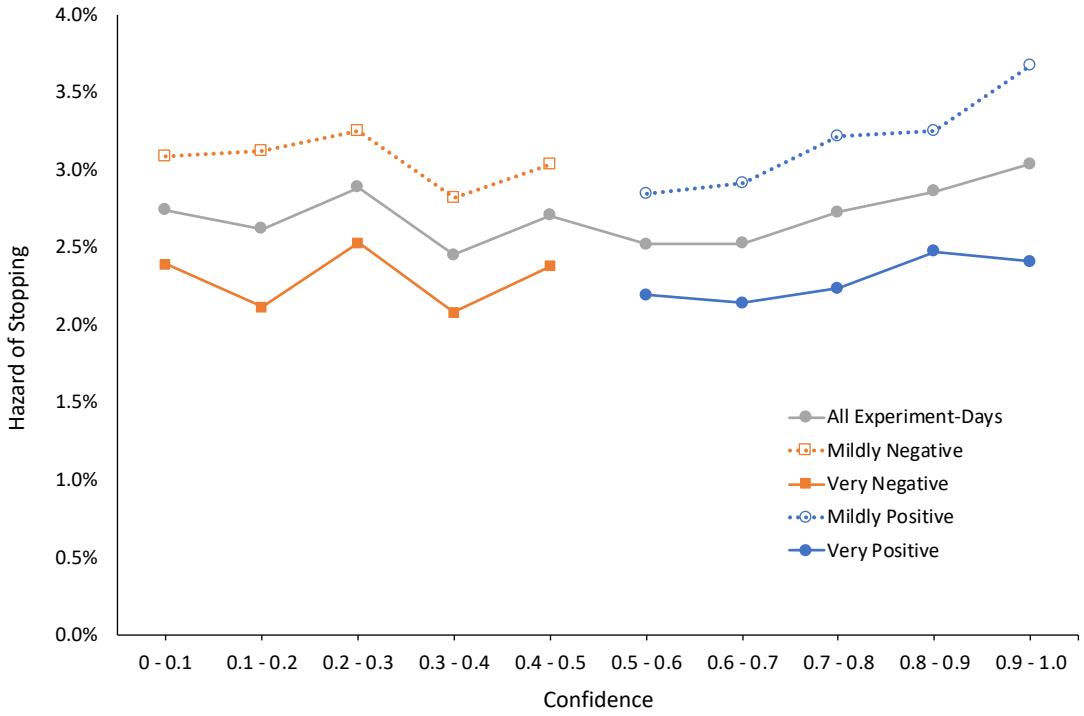
The model free analysis just presented shows that stopping experiments on a particular day is associated with how many days they have been running, the day of the week, the effect size,

Figure 5: Histograms of Experiments' Lift



Lift values between -0.6 and 0.6 constitute 97.3% (left) and 99.3% (right) of experiment-days.

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



and the confidence level or p-value achieved. The association between stopping and the p-value might reflect p-hacking. However, it may also result from experimenters being exceedingly good at setting the pre-determined sample size or run-time to detect the experimental effect. 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 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 indicate a U-shaped relation between confidence and stopping, mirroring the pattern in Figure 6. 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. However, the pattern in Model (2) including the control variables is 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 evidence 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 U-shaped relation between confidence and the propensity to stop is robust, and so is the sign reversal of the interaction between confidence and lift.

4 Do experimenters p-hack?

Since our data are observational, the model-free and model-based results reported so far do not establish causality. Consequently, they do not document p-hacking. 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 up" at critical confidence levels.

The critical levels we use are those typically used for null-hypothesis testing: 1%, 5% and 10% on the low end (where the baseline outperforms the best variation), and 90%, 95% and 99% on the high end (where the best variation outperforms the baseline). Because we have access to the actual confidence values, but users of the platforms only see values rounded to 2-digits, the actual critical levels we use are: 1.5%, 5.5%, 10.5%, 89.5%, 94.5% and 98.5%. For example, to assess whether people p-hack when the platform announces that their experiment has achieved 90% confidence, we use the 89.5% threshold.

In small windows around these values, 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

Table 4: Hazard Regression Results

	(1)	(2)
0–0.1	-3.3453^{***}	-4.3611^{***}
	(0.0000)	(0.0000)
0.1–0.2	-3.4056^{***}	-4.3498^{***}
	(0.0000)	(0.0000)
0.2–0.3	-3.3292^{***}	-4.2563^{***}
	(0.0000)	(0.0000)
0.3–0.4	-3.4927^{***}	-4.4421^{***}
	(0.0000)	(0.0000)
0.4–0.5	-3.4212^{***}	-4.3922^{***}
	(0.0000)	(0.0000)
0.5–0.6	-3.5222^{***}	-4.4606^{***}
	(0.0000)	(0.0000)
0.6–0.7	-3.4880^{***}	-4.4327^{***}
	(0.0000)	(0.0000)
0.7–0.8	-3.3768^{***}	-4.3384^{***}
	(0.0000)	(0.0000)
0.8–0.9	-3.3572^{***}	-4.3303^{***}
	(0.0000)	(0.0000)
0.9–1	-3.3230^{***}	-4.3195^{***}
	(0.0000)	(0.0000)
0–0.1× lift	0.1168^{*}	0.0304
	(0.0638)	(0.6630)
0.1–0.2× lift	0.2562^{***}	0.1730^{*}
	(0.0041)	(0.0658)
0.2–0.3× lift	0.1928^{**}	0.1035
	(0.0274)	(0.2586)
0.3–0.4× lift	0.1966^{**}	0.1074
	(0.0359)	(0.2733)
0.4–0.5× lift	0.2285^{***}	0.1845^{**}
	(0.0095)	(0.0449)
0.5–0.6× lift	-0.1584^{*}	−0.1113
	(0.0645)	(0.2128)
0.6–0.7× lift	-0.2322^{***}	-0.1652^{*}
	(0.0081)	(0.0720)
0.7–0.8× lift	-0.2860^{***}	-0.2137^{**}
	(0.0004)	(0.0125)
0.8–0.9× lift	-0.2148^{***}	-0.1559^{**}
	(0.0043)	(0.0497)
0.9–1× 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. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

effect size and sample size, both of which are driven by the random process of visitor arrival to the site.

4.1 RDD Hazard Model and Data Windows

We follow the standard 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).[3] Consequently, we occasionally observe the same experiment both below and above the cutoff.

We limit the data to the experiment-day observations that fall within a specific window, and use two logistic specifications for the hazard that i stops experiment j on day t.[4] The first includes only linear effects of confidence:

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

and the second includes linear and quadratic effects:

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

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. Following Gelman and Zelizer (2015) and Gelman and Imbens (2017), we do not consider higher order polynomials. In both models β_D captures the size of the discontinuity in the logit-hazard of stopping at the critical threshold. As such, positive values of β_D represent the causal effect of reaching a p-value threshold on stopping, and hence the existence of p-hacking as defined in this study.

Since there are no compelling rules for choosing the width of the window around the suspected discontinuity, we use a variety of widths to assess robustness as recommended in the literature. Specifically, we use 15 window widths ranging from ± 0.001 to ± 0.015. Choosing between wide and

[3] 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 that analyze only terminal effect sizes, standard errors and p-values.

[4] This 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).

narrow windows involves a trade-off . Wider windows contain more observations and hence typically yield higher power, whereas narrower windows typically offer better randomization. The linear specification is better suited for smaller windows in which the linearity assumption is more likely to be valid and that contain a relatively small number of observations. The quadratic specification is better suited for wider windows where the reverse holds.

Again, the key assumption for causal identification is that the continuous confidence variable behaves randomly within the narrow window around the critical level. Such randomness implies that covariates shown in the previous Section to be associated with stopping behavior do not exhibit a trend or a discrete jump at the critical confidence levels. Figure 7 presents histograms of these covariates for the 89.5% and 94.5% critical levels binned by the level of confidence. There are no pronounced jumps at the critical levels or trends throughout the windows, except possibly for the sample size in the 94.5% window. The latter may reflect the direct mathematical relation between confidence, effect size and sample size. This caveat does not apply to the 89.5% confidence window.

4.2 RDD Results

This Section presents evidence of p-hacking around 90% and 95% confidence (as announced by the platform to users), using both RDD tables and discontinuity plots. Evidence of p-hacking at the other four levels is discussed only briefly afterwards.

Table 5 presents the β_D estimates capturing discontinuities at 90%, for 15 windows and both the linear and quadratic specifications. The linear model detects discontinuities, i.e., p-hacking, in three contiguous narrow windows (0.005 – 0.007). The quadratic model detects discontinuities in six wider contiguous windows (0.008 – 0.013). This pattern is consistent with the advantages and disadvantages of wider windows containing more data but also greater deviations from linearity. Table 5 shows that discontinuities are detected by the linear but not the quadratic model when the former fits better after AIC penalization for additional parameters. In other words, discontinuities are detected by the linear but not the quadratic model in the absence of non-linearities. Conversely, discontinuities are detected by the quadratic model but not the linear model when the former fits about the same or better, reflecting the presence of non-linearities. Exponentiating the β_D estimates of the quadratic model in the 0.008 – 0.013 windows, shows that the effects are fairly large: at the

Figure 7: 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

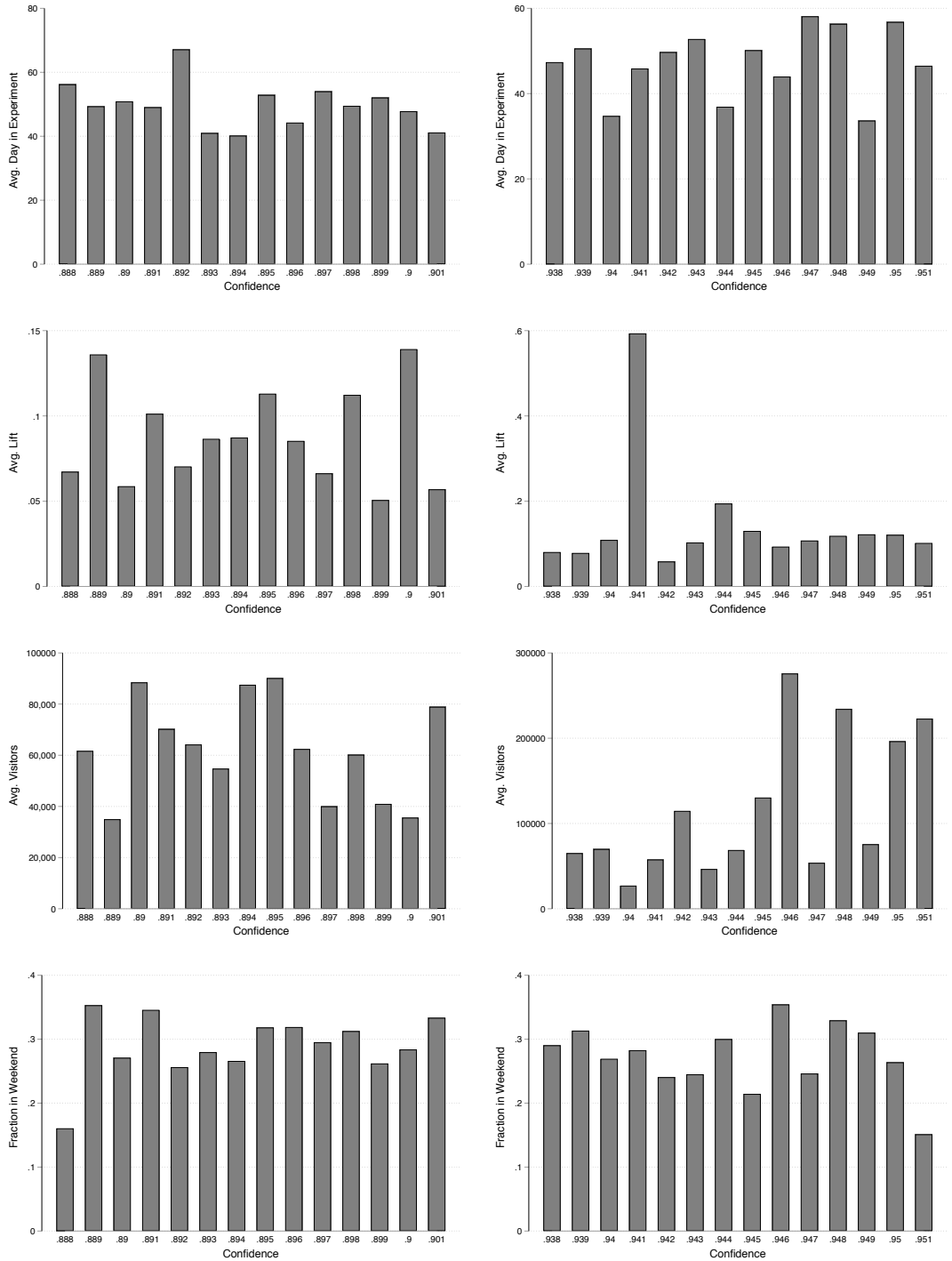


Table 5: Regression discontinuities at 89.5% confidence (reported as 90% to users)

		Linear Specification		Quadratic Specification		Relative Model Fit
Window Width	N	β_D	p-value	β_D	p-value	AIC diff (Quad-Lin)
.001	145	−2.142	0.2916			
.002	307	1.735	0.4154	−2.610	0.4379	−1.207
.003	450	1.844	0.1768	1.526	0.4456	3.949
.004	611	1.192	0.2965	1.786	0.2964	3.556
.005	791	1.633*	0.0891	0.803	0.5599	3.276
.006	950	1.364*	0.0880	1.546	0.2144	3.897
.007	1,133	1.401**	0.0446	1.316	0.2418	2.869
.008	1,290	0.849	0.1878	2.074**	0.0449	1.322
.009	1,461	0.759	0.2087	1.881*	0.0588	0.596
.010	1,663	0.183	0.7476	2.082**	0.0217	−4.709
.011	1,839	−0.143	0.7945	2.043**	0.0169	−9.05
.012	2,010	−0.056	0.9145	1.629*	0.0501	−4.163
.013	2,155	−0.169	0.7374	1.509*	0.0598	−4.9
.014	2,305	−0.231	0.6391	1.079	0.1420	−1.945
.015	2,464	−0.181	0.7046	0.803	0.2481	0.211

N = number of experiment-days in the window.

β_D = change at the critical level in the log-odds of stopping. p-values are in parentheses.

Empty cells denote lack of convergence due to insufficient observations.

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

discontinuity, the odds of termination jump up by a factor between 5 and 8.

Table 6 presents similar evidence of p-hacking at 95% confidence, be it in 5 rather than 9 windows. The linear model detects a discontinuity in one very narrow window, whereas the quadratic model detects discontinuities in the next four windows. As before, the discontinuity is detected by the linear model when that model has an AIC that is at least 2.8 larger than the quadratic model, and exponentiating the significant estimates of β_D indicates large effect sizes.

As recommended by Gelman and Zelizer (2015), we complement the RDD regressions with discontinuity plots showing data and the fitted model. In each window, 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. Figure 8 shows plots for four of the discontinuity estimates. Circles indicate the empirical hazard. A circle with zero value indicates no stops within that bin. Lines indicate the predicted value of a linear probability model on observations on each side of the window. The top-left graph provides visual confirmation of the discontinuity at 89.5% detected by the linear model in the narrow range of ±0.7%. The top right provides visual confirmation of the same discontinuity detected by the quadratic model in the wider range of ±1.1%. The bottom two graphs provide similar visual confirmation of the discontinuity detected at 94.5%. Notably, the jump in probabilities at the discontinuities map tightly into the estimates in Tables 5 and 6.

Table 6: Regression discontinuities at 94.5% confidence (reported as 95% to users)

		Linear Specification		Quadratic Specification		Relative Model Fit
Window Width	N	β_D	p-value	β_D	p-value	AIC diff (Quad-Lin)
.001	178	2.179	0.3255			
.002	372	2.723**	0.0470	2.946	0.2085	3.936
.003	516	1.386	0.1888	3.898*	0.0518	−0.196
.004	690	1.212	0.1639	2.502*	0.0802	1.859
.005	868	0.762	0.3544	2.741**	0.0496	−0.321
.006	1042	0.428	0.5538	2.069*	0.0688	−0.758
.007	1221	0.255	0.7106	1.559	0.1237	−0.404
.008	1401	0.312	0.6193	1.087	0.2634	2.71
.009	1578	0.484	0.4220	0.707	0.4325	3.837
.010	1772	0.354	0.5352	0.769	0.3692	3.525
.011	1936	0.341	0.5307	0.693	0.3858	3.61
.012	2125	0.236	0.6530	0.793	0.3042	2.977
.013	2290	0.054	0.9152	0.827	0.2597	1.378
.014	2489	0.195	0.6865	0.485	0.4794	2.905
.015	2699	0.273	0.5572	0.355	0.5934	3.527

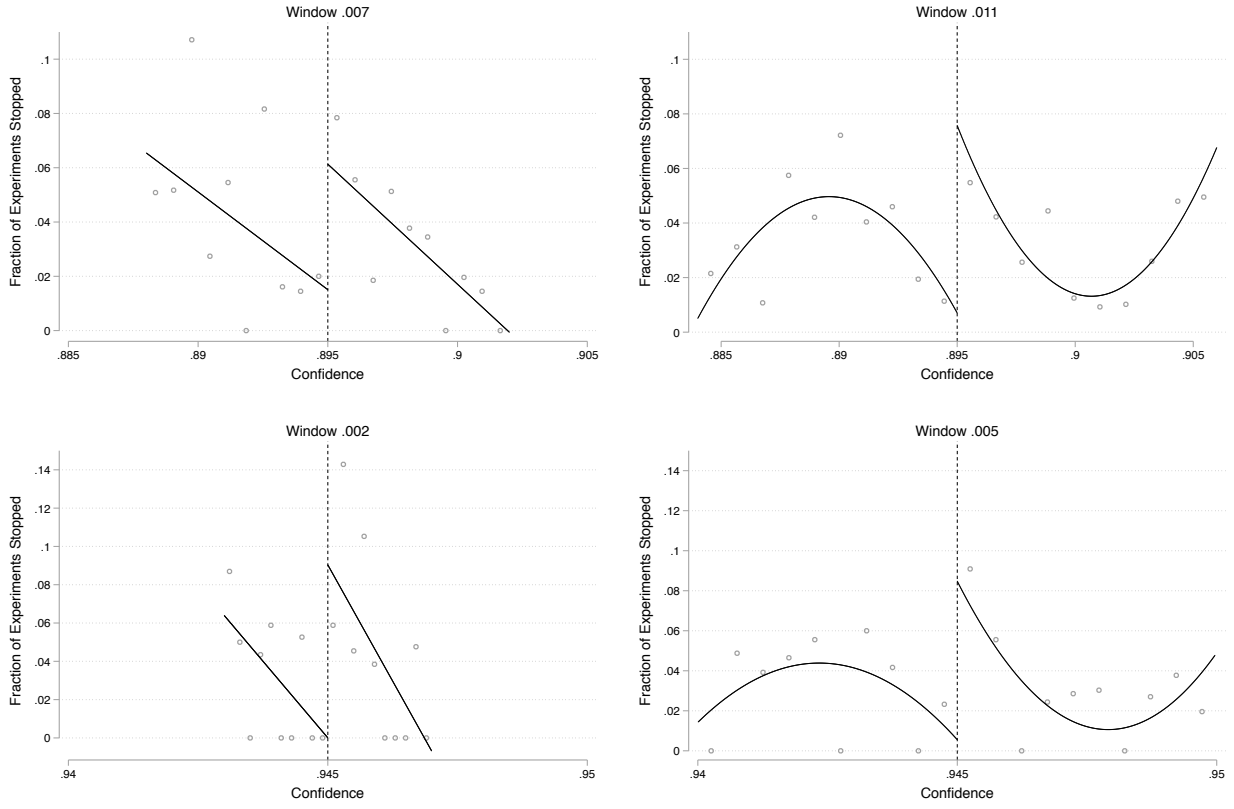
N = number of experiment-days in the window.

β_D = change at the critical level in the log-odds of stopping. p-values are in parentheses.

Empty cells denote lack of convergence due to insufficient observations.

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

Figure 8: Regression discontinuity plots of stopping at of 89.5% (top) and 94.5% (bottom)



We also conducted RDD hazard analyses, using the same 15 windows, for p-hacking at 1.5%, 5.5%, 10.5% and 98.5% (i.e., 1%, 5%, 10% and 99% as shown to users). We find evidence of p-hacking at 1% as shown to the user, but not the other three levels. Regression tables similar to Tables 5 and 6, and histograms similar to those in Figure 7, are presented in the Online Appendix.

In summary, there is evidence of p-hacking for both positive and negative lifts. When facing positive lifts, people p-hack at the more liberal of the standard levels (90%, 95%), and when facing negative lifts they p-hack only at the most stringent of the standard levels (1%). Assuming that the baseline is the status quo, such that a positive lift is a success and a negative lift is a failure, the data indicate that people p-hack when facing both good or bad news, but that they are much quicker to "pull the p-hacking trigger" when facing good news rather than bad news.

4.3 What fraction of experimenters p-hack?

The discontinuities documented so far pertain 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 where we find the most compelling evidence of p-hacking (Table 5).

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 $\beta_{DX^2} = 0$:

$$v^{NP}_{ijt} = \beta^{NP}_0 + \beta^{NP}_X \cdot X_{ijt} + \beta^{NP}_{X^2} \cdot X^2_{ijt}$$

Given the need for a large number of observations when estimating latent class models and the pattern in Table 5, we perform this analysis using the quadratic specification and a window width of $\pm 1.1\%$. This results in $1,839$ experiment-days from 373 of the original 916 experimenters.

The likelihood of the data 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 1.69 [95% CI: 0.02, 3.37], and the fraction of p-hackers w to be 57.06% [95% CI: 51.49%, 62.47%]. Hence, when experiments are within the $89.5\% \pm 1.1\%$ confidence window, slightly more than half the experimenters p-hack.

We also estimated a variant of the model where we asses whether the experience and industry of the experimenter is associated with being a p-hacker. We find that p-hacking is more likely in the Media industry and less likely in the High-Tech industry (odds ratios of 4.23 and 0.42 respectively; both $p < 0.05$), but unrelated to experience on the platform (Table A6 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 7, 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 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 7: 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 $\pi_0 = \frac{m_0}{m}$ 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 at either the 90%/10% or the 95%/5% critical levels.[5] $\mathbb{E}(F)$ is estimated as 10% or 20% of $\pi_0 \cdot m$ where m is simply the number of experiments in the data, and π_0 is estimated using the Storey and Tibshirani (2003) method. 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 compute confidence intervals for the estimated FDRs using bootstrapping with 1,000 samples.

Table 8 presents the estimates of π_0 and the implied FDR for the 90%/10% and the 95%/5% 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, about 38% of the results significant at 90%/10% are true nulls, and about 27% of those significant at 95%/5% are true nulls.

Table 8: Fraction of True Null Effects and FDRs

Discovery Threshold	$\widehat{\pi}_0$	$\widehat{\text{FDR}}$	FDR 95% C.I.
90%/10%	72.51%	38.08%	[32.33%, 43.82%]
95%/5%	72.65%	26.94%	[23.01%, 30.87%]

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 acts 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, the counterfactual FDRs reflect the

[5] 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.

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} controls for this difference in sample size.

To estimate the causal impact of p-hacking 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 57.06% (Section 4.3).[6] The FDR if 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 9 presents the actual and counterfactual FDR estimates, and the increase in FDR due to p-hacking using the previous estimate $\widehat{w} = 57.06\%$. 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 42%.

Table 9: 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%	38.08%	35.94%	32.89%	3.75%	9.09%
95%/5%	26.94%	25.6%	23.39%	2.34%	6.22%

[6] When using $w = 57.06\%$, we assume that the fraction of p-hackers in the full sample of 916 experimenters is the same as the fraction of p-hackers among the 373 experimenters with experiments in the $89.5\% \pm 1.1\%$ confidence window.

6 What is the cost of inflating the FDR?

p-Hacking increased the average FDR among p-hacked experiments from 33% to 42%. This generates two possible costs for a company engaging in p-hacking. 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, one 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 10. 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. 72.5%). 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. Since 55.9% of the experiments have $\widehat{\theta} > 0$ and the values in Table 10 imply that $\mathbb{E}[\theta|\widehat{\theta} > 0] = 0.0232$, 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 10: 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 whether online A/B experimenters p-hack by stopping their experiments based on the p-value of the treatment effect. We use data on 2,101 experiments and a regression discontinuity design to detect p-hacking.

We find evidence of p-hacking at 90% and 95% confidence for positive effects and at 1% for negative effects. The evidence is the strongest at 90%. About 57% of experimenters p-hack when their experiments are within the 89.5%±1.1% confidence window. Notably, these findings come from a platform that declared winners at 95% confidence for positive effects and losers at 5% confidence for negative effects. Hence, p-hacking cannot simply be attributed to the recommendation of the platform.

We also investigate the consequences of p-hacking on the false discovery rate (FDR). We find that in roughly 73% of the experiments the effect is truly null, and that the FDR is 38% at the 90%/10% critical levels, and 27% at the 95%/5% 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 FDR among all experiments from 33% to 38%, and the average FDR among p-hacked experiments from 33% to 42%. Assuming that following such a false discovery experimenters stop searching for more effective treatments, the expected cost of a false discovery is estimated as a forgone gain of 1.95% in lift. This corresponds to the 76^{th} percentile of observed lifts.

The behavior of experimenters in our data seems 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. If, in contrast, experiments are run to maximize 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.

Finally, on a more positive note, we find that stopping an experiment early or late is not driven solely by p-hacking. Specifically, we find a pronounced day-of-the-week pattern, a 7-day cycle in the first 35 days, and a tendency to terminate sooner when observing effects small rather than large in magnitude.

Our findings add urgency to earlier calls for research on experimenters' intentions and decision

rules when running and analyzing experiments (e.g., Leek et al. 2017 and Greenland 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 an additional facet of p-hacking that has received little attention in previous discussions of the practice of statistical inference. A/B testers often act as agents, either as employees or as third-party service providers. Consequently they may have strong incentives to show significant results, preferably positive. Our finding that experimenters from the Media industry are more likely to p-hack gives some empirical justification to that concern. We hope that future studies in which the agency relationship of the experimenter is known will be able to provide empirical insights on this issue.

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%), our study puts forth a specific explanation for this by estimating that 73% of the A/B test effects in our sample of 2,101 experiments are truly null. This supports earlier suspicions that some of the frustrations with A/B testing may stem from the ineffectiveness of the interventions being tested rather than the tool itself (Fung 2014).

There are at least four strategies to address p-hacking through data peeking and 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 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, p. 785). 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, e.g., using multi-armed bandits (Schwartz et al. 2017). We hope that our findings will provide further impetus to develop strategies accommodating p-hacking and to study their effectiveness.

References

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.

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.

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 (2017). 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.

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.

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.

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.

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.

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.

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

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

Table A2: Regression discontinuities at 1.5% confidence (reported as 1% to users)

		Linear Specification		Quadratic Specification		Relative Model Fit
Window Width	N	β_D	p-value	β_D	p-value	AIC diff (Quad-Lin)
.001	80					
.002	322	4.876	0.4192	2.575	0.8433	3.576
.003	474	1.216	0.5330	3.771	0.5093	2.916
.004	646	1.864	0.2701	4.854	0.4008	−2.834
.005	815	1.422	0.2403	1.590	0.4806	2.844
.006	1019	0.979	0.3400	2.174	0.2441	3.255
.007	1197	1.315	0.1553	1.270	0.4115	3.666
.008	1418	1.376*	0.0869	1.301	0.3763	1.54
.009	1642	1.600**	0.0346	0.981	0.4301	2.824
.010	1923	1.226*	0.0742	1.619	0.1749	3.604
.011	2255	0.846	0.1678	2.210*	0.0763	−0.407
.012	2691	0.726	0.2035	2.429*	0.0549	−5.388
.013	3049	0.913*	0.0966	1.318	0.1855	0.775
.014	3518	0.691	0.1832	1.692*	0.0863	−1.107
.015	7852	0.349	0.4771	1.889**	0.0386	−1.62

N = number of experiment-days in the window.

β_D = change at the critical level in the log-odds of stopping. p-values are in parentheses.

Empty cells denote lack of convergence due to insufficient observations.

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

Table A3: Regression discontinuities at 5.5% confidence (reported as 5% to users)

		Linear Specification		Quadratic Specification		Relative Model Fit
Window Width	N	β_D	p-value	β_D	p-value	AIC diff (Quad-Lin)
.001	141	−11.639	0.5006			
.002	271	−0.870	0.7022	−2.570	0.5534	3.666
.003	411	−1.419	0.3608	−1.000	0.7125	3.649
.004	537	−1.296	0.3271	−0.870	0.7321	1.416
.005	692	−0.354	0.7292	−1.849	0.3236	2.681
.006	825	−0.281	0.7578	−1.730	0.3564	0.914
.007	974	−0.450	0.5952	−0.839	0.5810	1.768
.008	1117	−0.489	0.5519	−0.551	0.6430	3.882
.009	1273	−0.817	0.2967	−0.219	0.8437	3.351
.010	1397	−0.790	0.2871	−0.385	0.7215	3.676
.011	1527	−0.885	0.2166	−0.349	0.7427	3.563
.012	1657	−0.868	0.1996	−0.546	0.5944	3.824
.013	1788	−0.669	0.3099	−0.836	0.3770	2.469
.014	1919	−0.643	0.3125	−0.811	0.3950	3.506
.015	2014	−0.267	0.6633	−1.272	0.1806	1.237

N = number of experiment-days in the window.

β_D = change at the critical level in the log-odds of stopping. p-values are in parentheses.

Empty cells denote lack of convergence due to insufficient observations.

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

Table A4: Regression discontinuities at 10.5% confidence (reported as 10% to users)

		Linear Specification		Quadratic Specification		Relative Model Fit
Window Width	N	β_D	p-value	β_D	p-value	AIC diff (Quad-Lin)
.001	50					
.002	221	1.752	0.3775	−4.807	0.8484	−4.701
.003	352	0.269	0.8502	13.597	0.3154	−2.586
.004	475	0.476	0.7047	1.135	0.6229	2.942
.005	598	−0.012	0.9915	1.570	0.4681	2.414
.006	725	0.387	0.7081	0.211	0.9001	3.459
.007	846	−0.120	0.9019	0.838	0.5879	3.248
.008	944	−0.145	0.8749	0.681	0.6551	3.147
.009	1048	−0.025	0.9773	0.448	0.7673	2.698
.010	1157	0.048	0.9546	0.455	0.7586	2.342
.011	1264	0.790	0.3435	−0.508	0.6570	1.508
.012	1375	1.476*	0.0756	−0.779	0.4851	−3.159
.013	1468	1.250	0.1111	−0.250	0.8189	0.743
.014	1572	1.618**	0.0386	−0.192	0.8523	−1.254
.015	1690	1.496**	0.0442	0.330	0.7426	1.462

N = number of experiment-days in the window.

β_D = change at the critical level in the log-odds of stopping. p-values are in parentheses.

Empty cells denote lack of convergence due to insufficient observations.

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

Table A5: Regression discontinuities at 98.5% confidence (reported as 99% to users)

		Linear Specification		Quadratic Specification		Relative Model Fit
Window Width	N	β_D	p-value	β_D	p-value	AIC diff (Quad-Lin)
.001	238	−1.763	0.3055			
.002	451	−1.987	0.1034	−1.726	0.3219	3.274
.003	723	−0.646	0.4593	−3.082*	0.0763	−0.588
.004	950	−0.426	0.5875	−1.461	0.2353	2.655
.005	1183	0.246	0.7258	−1.716	0.1204	−2.301
.006	1440	0.207	0.7430	−0.902	0.3542	1.292
.007	1704	−0.010	0.9869	−0.287	0.7336	3.311
.008	2041	0.258	0.6497	−0.325	0.6749	0.911
.009	2308	0.420	0.4351	−0.301	0.6885	1.381
.010	2641	0.529	0.2904	−0.315	0.6682	1.725
.011	2928	0.439	0.3505	−0.062	0.9310	3.047
.012	3276	0.559	0.2130	−0.056	0.9332	2.538
.013	3729	0.390	0.3529	0.286	0.6670	3.271
.014	4319	0.260	0.5112	0.480	0.4630	2.155
.015	9404	−0.006	0.9875	1.000	0.1200	−0.49

N = number of experiment-days in the window.

β_D = change at the critical level in the log-odds of stopping. p-values are in parentheses.

Empty cells denote lack of convergence due to insufficient observations.

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

Table A6: Latent class model estimates

	Model (1)		Model (2)	
Response equations	non p-hacking	p-hacking	non p-hacking	p-hacking
Intercept	−26.218	−3.163***	−26.164*	−3.038***
	(17.118)	(0.750)	(15.872)	(0.834)
X	−7822.653	−304.089	−7814.089	−282.071
	(5577.704)	(309.629)	(5175.509)	(314.684)
X^2	−652068.805	−31239.980	−650250.336	−29868.347
	(458185.194)	(27589.705)	(425797.003)	(28037.181)
D		1.694**		1.687**
		(0.855)		(0.856)
$D \cdot X$		−358.738		−401.411
		(389.094)		(394.130)
$D \cdot X^2$		88993.826***		89252.380***
		(34258.024)		(34411.933)
Class membership equation				
Intercept	−0.285**		−0.121	
	(0.115)		(0.195)	
High-Tech			0.865***	
			(0.216)	
Media			−1.44**	
			(0.561)	
Professional Services			0.356	
			(0.413)	
Retail			−0.464	
			(0.299)	
log(Past # Experiments)			−0.086	
			(0.068)	
LL	-259.43		-256.82	

Standard Errors in Parentheses. $N = 1839$. * $p < 0.1$ ** $p < 0.05$ *** $p < 0.01$.

Note: the class membership equation models the probability of not being a p-hacker.

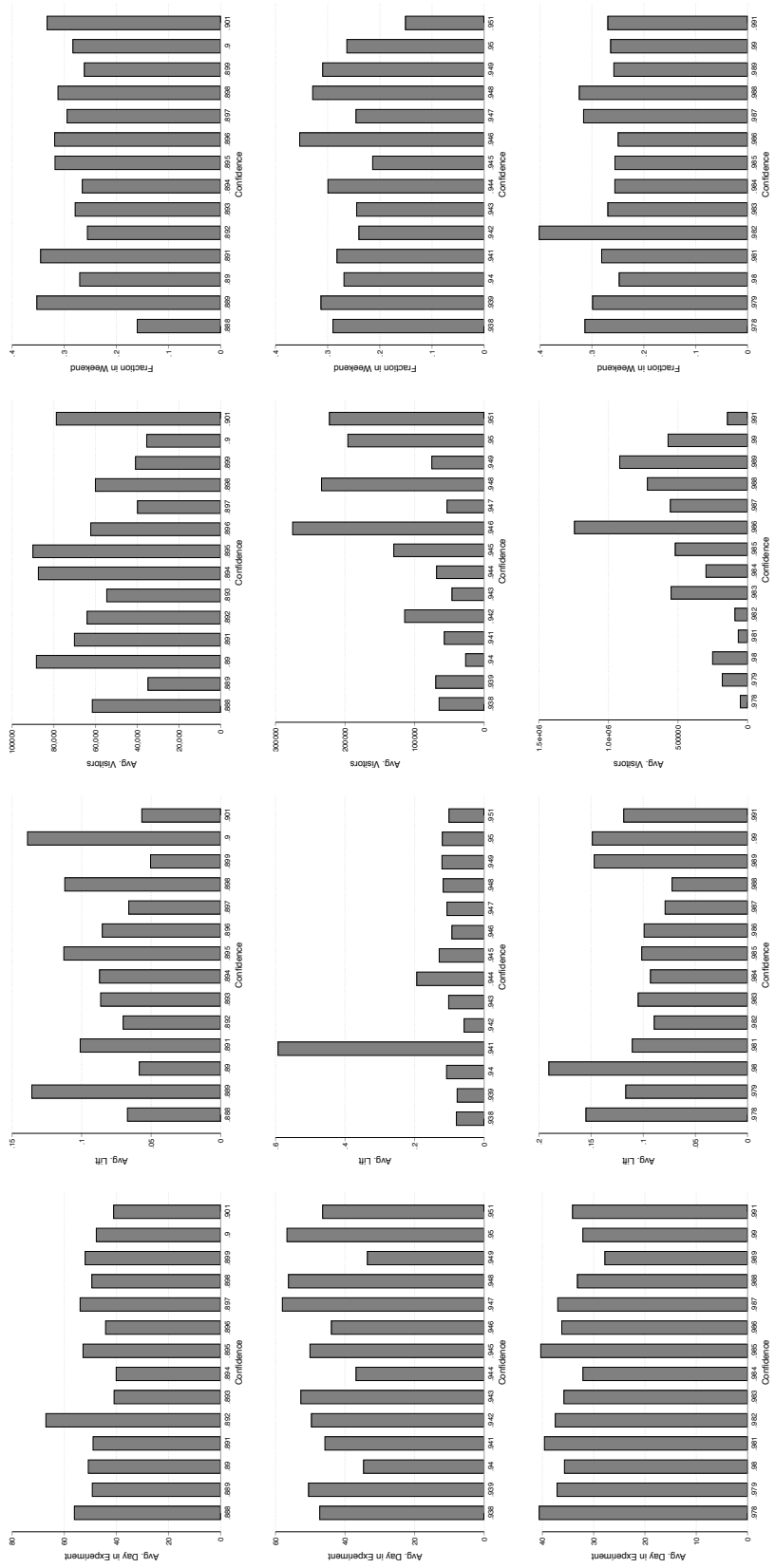


Figure A1: Histograms of average values of covariates inside 0.001 confidence-wide bins around different critical values of confidence. Critical Values: First row .895, Second row .945, Third row .985. Covariates (from left to right): Day in the experiment, Lift, Number of Visitors, Weekend Indicator

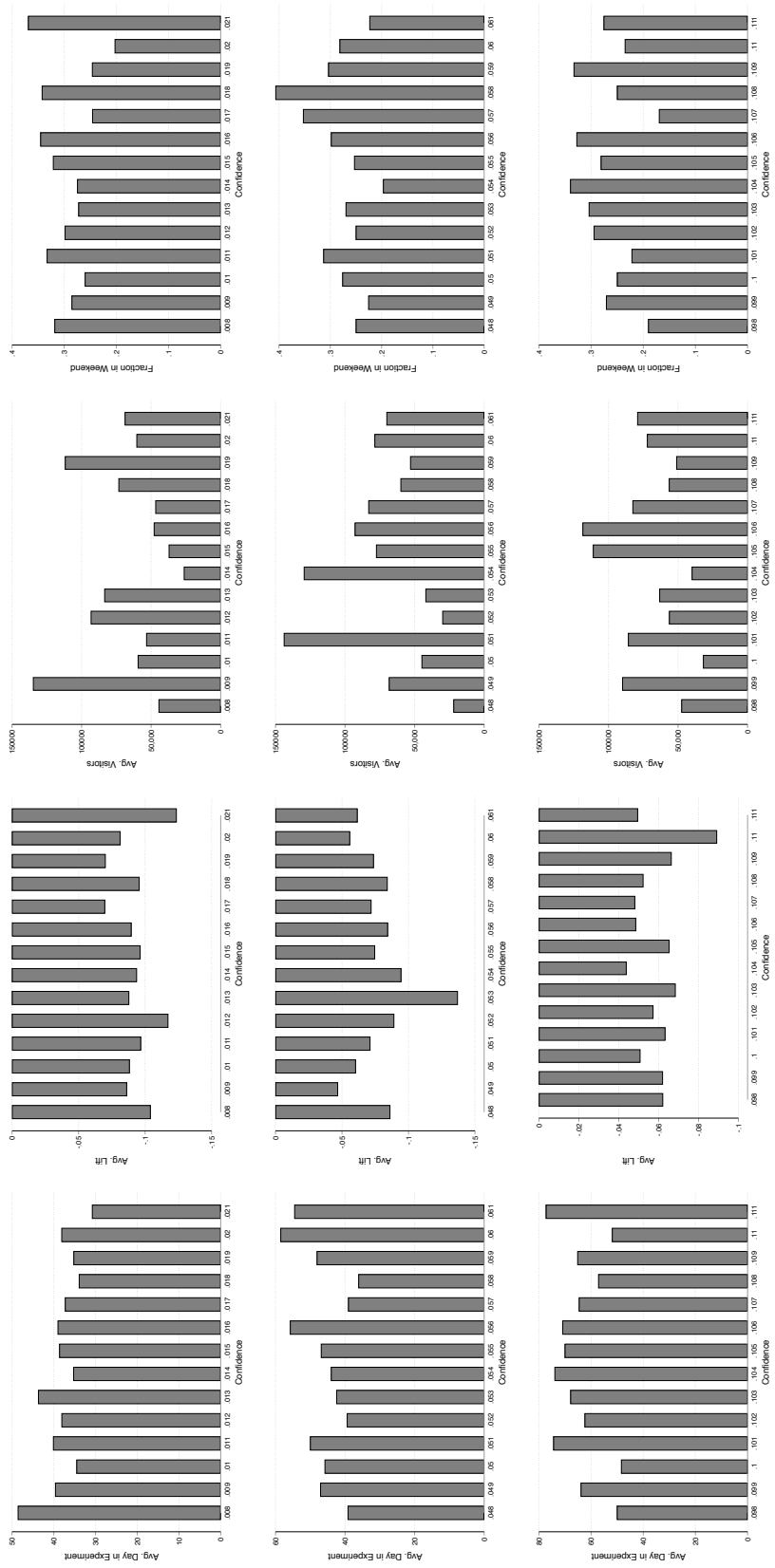


Figure A2: Histograms of average values of covariates inside 0.001 confidence-wide bins around different critical values of confidence. Critical Values: First row .015, Second row .055, Third row .105. Covariates (from left to right): Day in the Experiment, Lift, Number of Visitors, Weekend Indicator

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 $\pi_0 = 0$:

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