This article was downloaded by: [128.91.108.209] On: 11 April 2022, At: 09:15
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

False Discovery in A/B Testing

Ron Berman, Christophe Van den Bulte

To cite this article:
Ron Berman, Christophe Van den Bulte (2021) False Discovery in A/B Testing. Management Science

Published online in Articles in Advance 30 Dec 2021

. https://doi.org/10.1287/mnsc.2021.4207

Full terms and conditions of use: https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2021, INFORMS

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

informs
http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–21
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

False Discovery in A/B Testing

Ron Berman,[a] **Christophe Van den Bulte**[a]

[a] Marketing, The Wharton School of the University of Pennsylvania, Philadelphia, Pennsylvania 19104

Contact: ronber@wharton.upenn.edu, https://orcid.org/0000-0002-8594-3627 (RB); vdbulte@wharton.upenn.edu, https://orcid.org/0000-0001-9708-1596 (CVdB)

Received: March 10, 2020
Revised: March 15, 2021; May 24, 2021
Accepted: June 2, 2021
Published Online in Articles in Advance: December 30, 2021

https://doi.org/10.1287/mnsc.2021.4207

Copyright: © 2021 INFORMS

Abstract. We investigate what fraction of all significant results in website A/B testing is actually null effects (i.e., the false discovery rate (FDR)). Our data consist of 4,964 effects from 2,766 experiments conducted on a commercial A/B testing platform. Using three different methods, we find that the FDR ranges between 28% and 37% for tests conducted at 10% significance and between 18% and 25% for tests at 5% significance (two sided). These high FDRs stem mostly from the high fraction of true null effects, about 70%, rather than from low power. Using our estimates, we also assess the potential of various A/B test designs to reduce the FDR. The two main implications are that decision makers should expect one in five interventions achieving significance at 5% confidence to be ineffective when deployed in the field and that analysts should consider using two-stage designs with multiple variations rather than basic A/B tests.

History: Accepted by Eric Anderson, marketing.
Funding: The authors acknowledge funding by the Wharton Dean's Research Fund.
Supplemental Material: The data files and online appendix are available at https://doi.org/10.1287/mnsc.2021.4207.

Keywords: statistics • design of experiments • decision analysis • inference • A/B testing • false discovery rate

1. Introduction

Marketers increasingly use online experiments (A/B tests) to inform their decisions. Such experimentation is facilitated by various A/B testing platforms like Adobe Target, Google Optimize, Monetate, Optimizely, and VWO. These platforms make it easy to randomly allocate users to treatment conditions and to measure their responses.

Despite the increasing popularity of website A/B testing, practitioners using it are often disappointed with the results. First, the great majority of effects are very small and statistically nonsignificant.[1] The same has been observed in digital advertising experiments (Blake et al. 2015, Lewis and Rao 2015, Johnson et al. 2017a, Gordon et al. 2019). Second, even when the intervention exhibits a statistically significant (or significant, for short) uplift, deploying it often generates no notable improvement in the field (Goodson 2014). In other words, the result does not replicate, implying that the original test result was a false positive or false discovery.

This study investigates false discovery in A/B testing by analyzing data from nearly 5,000 effects tested in 2,766 experiments run on Optimizely, the largest online A/B testing platform with roughly 35% of market share.[2] Specifically, we answer three questions. (i) How prevalent are false discoveries? (ii) To what extent does this prevalence stem from a high fraction of true nulls versus low power? (iii) What can firms do to improve their false discovery rate (FDR)? The answers to these questions not only quantify various facets of the false discovery problem affecting A/B testing but also point to promising ways to address it. In the process, we also explore a few additional questions, such as whether the FDR varies systematically across industries and experimenters' experience.

Our study provides three main insights. First, false discoveries are indeed quite prevalent in website A/B testing. Of all effects displaying statistical significance at 5%, about one in five are truly null. At 10% significance, the FDR is about one in three. Possible malpractice in data analysis, such as not accounting for multiple comparisons, will produce even higher FDRs.

Second, the disappointingly high FDR stems mostly from a high fraction of true nulls rather than high type I or type II error rates. Specifically, the main culprit is that true nulls account for about 70% of all the effects being tested. A similarly high fraction of null effects has been observed on Microsoft's Bing (Deng 2015), and our study generalizes this finding to a much greater set of experimenters, organizations, and industries. In contrast, inadequate power contributes only little to the high FDR. The average power in the experiments we analyze is 65%–70% at 10% significance, and the FDR of tests at that level of significance would still be 20% even if power was 100%. Neither would tightening the significance level fully resolve the high-FDR problem (e.g., the FDR remains 18%–25% at 5% significance and 5%–8% at 1% significance). For decision makers, these findings imply that possible disappointment with A/B testing stems not from deficiencies of the method itself but from the interventions being tested in the experiments.

Even so, our third insight pertains to improvements in the design of A/B tests that can reduce the FDR. Specifically, a simulation informed by our empirical estimates indicates that analysts should consider using two-stage designs with multiple variations rather than basic A/B tests. For parameter values representative of our sample of experiments, the FDR for one-sided tests at 5% significance improves from 24% to 12%.

We proceed as follows. Section 2 presents a formal definition of the FDR and its relation to the fraction of true nulls and the type I and II error rates. Section 2 also describes three methods to estimate the fraction of true nulls and the FDR. Section 3 shows how false discoveries not only cause unnecessary switching costs and disappointment with rolling out false discoveries with zero effect, but also lower the expected gains in effectiveness from running experiments. Section 4 describes the research setting and the data. Section 5 presents the main results, followed by Section 6 which documents various contingencies. Section 7 investigates how the design of A/B tests can be improved to lower the FDR, and Section 8 discusses to what extent the findings generalize beyond our specific research setting. Section 9 concludes with implications for decision makers and analysts.

2. False Discovery Rate

2.1. Definition

A basic A/B test is designed to assess the difference in outcomes of two versions of a web page. We call this difference the effect θ. The false discovery rate is the probability that a measured effect $\widehat{\theta}$ reflects a true null ($\theta = 0$), even though $\widehat{\theta}$ is statistically significant at some level of significance α. In this section, we define the FDR in mathematical terms and state its relation to the type I and II error rates.

Let θ be the true effect, and let its estimate $\widehat{\theta}$ be declared significant if the test statistic lies in a rejection region Γ. For simplicity of exposition, we assume that the test is based on a z score. For a two-sided test at $\alpha = 0.05$, $\Gamma = \{z : |z| \geq 1.96\}$. Considering a set of experiments where the true effect θ varies, the FDR is then $Pr(\theta = 0 \mid z \in \Gamma)$. In contrast, the type I error rate α is the probability that a significant measurement actually stems from a null, i.e., $\alpha = Pr(z \in \Gamma \mid \theta = 0)$. The type II error rate β is the probability that a nonsignificant result stems from a nonnull effect, i.e., $\beta = Pr(z \in \Gamma \mid \theta \neq 0)$. The power of the test or the probability that the measurement of a true nonnull effect is significant, $Pr(z \in \Gamma \mid \theta \neq 0)$, is simply $1-\beta$.[3] Finally, denote the probability that the true effect is null as $Pr(\theta = 0) = \pi_0$.[4] Using Bayes rule, we can express the FDR as

$$\begin{aligned} Pr(\theta = 0 \mid z \in \Gamma) &= \frac{Pr(z \in \Gamma | \theta = 0) Pr(\theta = 0)}{Pr(z \in \Gamma)} \\ &= \frac{Pr(z \in \Gamma | \theta = 0) Pr(\theta = 0)}{Pr(z \in \Gamma | \theta = 0) Pr(\theta = 0) + Pr(z \in \Gamma | \theta \neq 0) Pr(\theta \neq 0)} \\ &= \frac{\alpha \pi_0}{\alpha \pi_0 + (1-\beta)(1-\pi_0)}. \end{aligned} \qquad (1)$$

In a properly conducted and analyzed experiment, the FDR is a function of three elements: the probability that effects are truly null π_0, the type I error rate or significance level α, and the power $1-\beta$.

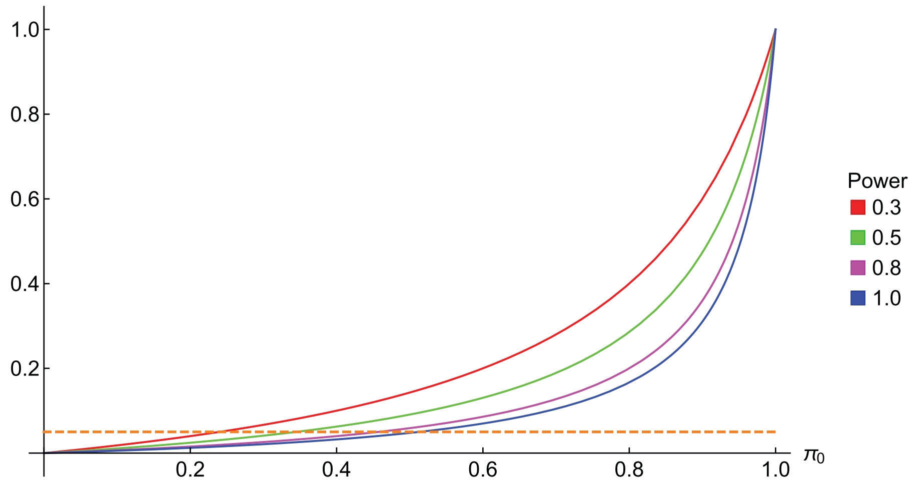
Note how the FDR differs from the type I error rate α. Whereas $\alpha = Pr(z \in \Gamma \mid \theta = 0)$, the FDR equals $Pr(\theta = 0 \mid z \in \Gamma)$ (i.e., the conditioning is reversed). Figure 1 displays the FDR as a function of π_0 and power $1-\beta$ for a fixed value of $\alpha = 0.05$. For many combinations of π_0 and power, the FDR is higher than α. For example, with $\alpha = 0.05$, $\pi_0 = 70\%$, and power at 80%, the FDR is 12.7%. Even with power at 100%, the FDR remains elevated at 10.4%.

Taking derivatives of Equation (1) confirms that the FDR increases with the fraction of true nulls π_0 and with the significance level α and decreases with the power $1-\beta$. Equation (1) also shows that any practices that increase α above its nominal level (e.g., 0.05) will increase the FDR at that level. Such practices include improperly testing multiple hypotheses and improperly testing hypotheses after peeking at the data (Benjamini and Hochberg 1995, Johari et al. 2017).

A second way to present the FDR and its relation to type I and II error levels is through the matrix shown in Table 1. It organizes measured effects (events) based on whether they stem from a true null or not and whether they are declared significant or not. The expected fraction of observations that are true nulls is π_0, and a fraction α of those will expectedly be declared significant. Of the expected $1-\pi_0$ fraction of nonnulls, a fraction $1-\beta$ will expectedly be declared significant. Hence, the expected fraction of all significant results that are false positives or false discoveries is the same as in Equation (1).

Finally, the FDR can also be defined starting from a mixture model. This has proven useful for quantifying the FDR empirically (e.g., Efron 2012). Let $f(z \mid \theta)$ be the probability density function (pdf) of the z scores conditional on the effect θ. Also, let $f_0(z) = f(z \mid \theta = 0)$, which is the standard Normal, and $f_1(z) = \int_{\theta \neq 0} f(z \mid \theta) h(\theta) d\theta$, where $h(\theta)$ is the pdf of the nonnull effects, such that

Figure 1. (Color online) How FDR at Multiple Power Levels Varies with π_0 When $\alpha = 0.05$



$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z). \tag{2}$$

Again, using Bayes rule, we can express the probability that a measured effect with a specific z score stems from the null as

$$\begin{aligned} Pr(\theta = 0|z) &= \frac{f(z|\theta = 0)Pr(\theta = 0)}{f(z)} \\ &= \frac{\pi_0 f_0(z)}{\pi_0 f_0(z) + (1 - \pi_0) f_1(z)}. \end{aligned} \tag{3}$$

Integrating the expression in (3) over the rejection region Γ that corresponds to the significance level α and its critical z score z^* results in

$$\begin{aligned} &Pr(\theta = 0 \,||\, z \,| \geq z^*) \\ &= \frac{\pi_0(1 - F_0(z^*) + F_0(-z^*))}{\pi_0(1 - F_0(z^*) + F_0(-z^*) + (1 - \pi_0)(1 - F_1(z^*) + F_1(-z^*))} \\ &= \frac{\pi_0 \alpha}{\pi_0 \alpha + (1 - \pi_0)(1 - \beta)}. \end{aligned} \tag{4}$$

2.2. Estimation

We estimate π_0 using three methods (A, B, and C) that use different inputs and modeling assumptions. Method A uses normal mixture modeling, an approach many researchers and data scientists are familiar with. Method B also uses normal mixture modeling but allows the estimation of the FDR, as first shown by Efron et al. (2001). It also allows the estimation of average power. The key idea underlying Method B has generated a multitude of specific implementations (e.g., Efron 2012, Scott et al. 2015), and the one we use extends Method A. Method C is a nonparametric approach originally developed by Storey (2002, 2003) and Storey and Tibshirani (2003), and it allows the estimation of π_0, the FDR, and average power.

Methods B and C use different aspects of the data to identify the value of π_0 used to estimate the FDR. Method B identifies π_0 from the presence of too large a spike in the middle and too many observations far into the tails of the distribution of z scores compared with the standard Normal when $\pi_0 = 1$. Method C identifies π_0 from the shape of the distribution of p-values compared with the uniform distribution when $\pi_0 = 1$. Other similarities and differences across the three methods are shown in Table 2. Together, they span various dimensions among the approaches currently used to compute the proportion of true nulls and the FDR (for a recent review, see Korthauer et al. 2019).

2.2.1. Method A.

Let $\widehat{\theta}_i$ be the estimated effect size from test i. We assume that the true effect comes from a mixture of nulls and nonnulls; the effect size θ_i is zero (null) with probability π_0 and $\theta_i \sim \mathcal{N}(\mu, \sigma^2)$ otherwise. Consistent with the central limit theorem, we assume that $\widehat{\theta}_i \sim \mathcal{N}(\theta_i, s^2)$. Here, σ represents the variation among the nonnull effects, and s represents the estimation error. The likelihood of an estimated

Table 1. True Nulls and Type I and II Error Rates

	Called significant (discovery)	Called not significant	Total
Null is true	$\alpha\pi_0$	$(1 - \alpha)\pi_0$	π_0
Alternative is true	$(1 - \beta)(1 - \pi_0)$	$\beta(1 - \pi_0)$	$1 - \pi_0$
Total	$\alpha\pi_0 + (1 - \beta)(1 - \pi_0)$	$(1 - \alpha)\pi_0 + \beta(1 - \pi_0)$	1

Table 2. Estimation Methods

	Method A	Method B	Method C
Inputs (DV*)	Effect size*	*z* score* and s.e.	*p*-value*
DV distribution under H_0	Normal	Standard Normal	Uniform
DV distribution under H_1	Normal (homoskedastic)	Normal (heteroskedastic)	Nonparametric
Estimation of π_0	Yes	Yes	Yes
Estimation of *FDR*	No	Yes	Yes
Estimation of power	No	Yes	Yes

Note. DV, dependent variable.

effect is

$$f(\widehat{\theta}_i) = \pi_0\phi(\widehat{\theta}_i;0,s^2) + (1-\pi_0)\phi(\widehat{\theta}_i;\mu,\sigma^2+s^2), \qquad (5)$$

where $\phi(x;\lambda,\tau^2)$ is the pdf of the normal distribution with mean λ and variance τ^2. From Equation (5), we estimate the four model parameters π_0, μ, σ, and s using maximum likelihood estimation (MLE).

2.2.2. Method B.

Method A assumes that s^2 is common across tests, regardless of sample size and the conversion rates in the A/B test. This does not take into account the different standard error (s.e.) used in each test. Method B addresses this limitation.

We incorporate the observed s.e. $\widehat{s}_i$ and assume $\widehat{\theta}_i \sim \mathcal{N}(\theta,\widehat{s}_i^2)$. The likelihood of $\widehat{\theta}_i$ conditional on $\widehat{s}_i$ is

$$f(\widehat{\theta}_i\,|\,\widehat{s}_i) = \pi_0\phi(\widehat{\theta}_i;0,\widehat{s}_i^2) + (1-\pi_0)\phi(\widehat{\theta}_i;\mu,\sigma^2+\widehat{s}_i^2). \qquad (6)$$

Equivalently, the test statistic (asymptotic z score) $z_i = \widehat{\theta}_i/\widehat{s}_i$ has the following conditional likelihood:

$$f(z_i\,|\,\widehat{s}_i) = \pi_0\phi(z_i;0,1) + (1-\pi_0)\phi(z_i;\mu/\widehat{s}_i,1+\sigma^2/\widehat{s}_i^2), \qquad (7)$$

which is the empirical analog of Equation (2). Note that the distribution of the nonnull z score is still Normal but with mean and variance decreasing in $\widehat{s}_i$ (Hung et al. 1997, Lu and Stephens 2019). Based on Equation (7), we estimate π_0, μ, and σ using MLE.

To estimate the FDR over the entire set of m test results, we plug the value of $\widehat{s}_i^2$ and the MLE estimates of π_0, μ, and σ^2 into Equation (4) where $\Phi(\cdot)$ is a normal cumulative distribution function (cdf),

$$\widehat{Pr}(\theta_i=0\,|\,\widehat{s}_i,\,|z|>z^*) = \frac{\widehat{\pi}_0\alpha}{\widehat{\pi}_0\alpha + (1-\widehat{\pi}_0)(1-\Phi(z^*;\widehat{\mu}/\widehat{s}_i,1+\widehat{\sigma}^2/\widehat{s}_i^2)+\Phi(-z^*;\widehat{\mu}/\widehat{s}_i,1+\widehat{\sigma}^2/\widehat{s}_i^2))} \qquad (8)$$

and integrate over the empirical distribution of $\widehat{s}_i$:

$$\widehat{FDR}(z^*) = \frac{\sum_{i=1}^{m}\widehat{Pr}(\theta_i=0\,|\,\widehat{s}_i,\,|z|>z^*)}{m}. \qquad (9)$$

2.2.3. Method C.

Both Methods A and B make the parametric assumption that the distribution of the nonnull effects is normal. Method C uses a nonparametric approach with the p-values as its input. The key idea is that under the null hypothesis, the p-values will be uniformly distributed, whereas under the alternative, the distribution of p-values will be skewed toward low values. Consequently, the fraction of effect sizes with very high p-values provides a good estimate of π_0.

To quantify the FDR, Method C uses Table 3, which is the empirical analog of Table 1. In the table, m, the total number of effects, and S, the number of significant effects, are both observed. In contrast, m_0, the number of true nulls, and F, the number of false positives, are not observed. As the number of estimated effects m increases, the fraction $\frac{F}{m_0}$ converges to the nominal significance level α, and the fraction $\frac{T}{m_1}$ converges to the power $1-\beta$. Also, $FDR = \mathbb{E}[\frac{F}{S}]$.[5]

As noted, the only observed values in Table 3 are m and S, but because $F = \alpha m_0 = \alpha\pi_0 m$ as m becomes very large, all we need to fill out the table and compute the FDR is an estimate of π_0 (Storey 2002, 2003; Storey and Tibshirani 2003). Specifically, we compute π_0 using the method proposed by Storey and Tibshirani (2003) and implement it using the default settings in the R package *q*value (Storey et al. 2019). The estimation consists of the following four steps.

1. Denote p_j as the p-value of effect j.
2. For a range of $\lambda = 0.05, 0.1, \ldots, 0.95$, calculate

$$\widehat{\pi}_0(\lambda) = \frac{\#\{\text{tests with } p_j > \lambda\}}{m(1-\lambda)}.$$

3. Fit the natural cubic spline $\widehat{g}$ of $\widehat{\pi}_0(\lambda)$ on λ.
4. Set the estimate of π_0 to be $\widehat{\pi}_0 = \widehat{g}(1)$.

Instead of using the fraction of observations with $p > 0.95$ or $p > 0.99$, this approach borrows strength from the entire distribution by fitting a flexible curve and taking the estimate at the limit of $\lambda = 1$. We use bootstrapping to compute confidence intervals (C.I.s) for $\widehat{\pi}_0$. To compute the FDR at a specific significance level α, we again follow Storey and Tibshirani (2003):

$$\widehat{FDR}(\alpha) = \frac{m\cdot\widehat{\pi}_0\cdot\alpha}{\#\{\text{tests with } p_j \le \alpha\}}. \qquad (10)$$

We calculate the FDR at the three levels of α most commonly used in the social sciences, 10%, 5%, and

Table 3. 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

1% (Leahey 2005, Brodeur et al. 2020), and "bookend" these with two additional levels, 20% and 0.1%.

3. The Cost of False Discoveries

Implementing false discoveries generates two kinds of costs. The first is a cost of omission: the gain foregone by deploying a false rather than true discovery. The second is a cost of commission: the costs incurred by deploying a false discovery rather than sticking with the current practice. We discuss each in turn.

Assume a decision maker runs an A/B experiment with one test and one control condition, where the control is the current implementation. The decision maker uses a one-sided test and switches to the treatment if the effect $\widehat{\theta}$ is positive and statistically significant at level α ($\widehat{\theta}/\widehat{s} > z_{1-\alpha}$).

The gain in effectiveness from such an experiment is zero if the observed effect is not significant; it is also zero if the observed effect is significant but the true effect θ is null, and it is non-zero with a value of θ otherwise. If the effects and the observed data are generated according to the normal mixture model introduced, then the expected gain (EG) in effectiveness for a given $\widehat{s}$ equals (see Online Appendix A)

$$EG = (1-\pi_0)\cdot\Phi\left(\frac{\mu - z_{1-\alpha}\cdot\widehat{s}}{\sqrt{\widehat{s}^2+\sigma^2}}\right)\cdot\left(\mu + \frac{\sigma^2}{\sqrt{\widehat{s}^2+\sigma^2}}\frac{\phi\left(\frac{\mu - z_{1-\alpha}\cdot\widehat{s}}{\sqrt{\widehat{s}^2+\sigma^2}}\right)}{\Phi\left(\frac{\mu - z_{1-\alpha}\cdot\widehat{s}}{\sqrt{\widehat{s}^2+\sigma^2}}\right)}\right) \tag{11}$$

$$= Pr(\theta \neq 0)\cdot Pr(\widehat{\theta}/\widehat{s} > z_{1-\alpha} | \theta \neq 0)\cdot E\left[\theta | \widehat{\theta}/\widehat{s} > z_{1-\alpha}, \theta \neq 0\right]. \tag{12}$$

These expected gains decrease with π_0 and increase with μ and σ because the latter drive larger true effects. The expected gains decrease with $\widehat{s}$ because lack of precision lowers the ability to detect and implement interventions with true effects larger than zero. More surprising is that the gains in effectiveness decrease in $z_{1-\alpha}$ when $z_{1-\alpha} \geq 0$ (i.e., the gains decrease as one tightens the significance level α used to declare discoveries). The reason is that fewer positive true effects are declared significant, i.e., $Pr(\widehat{\theta}/\widehat{s} > z_{1-\alpha} \mid \theta \neq 0)$ goes down, and this dominates the increase in the expected true effect conditional on significance $E[\theta \mid \widehat{\theta}/\widehat{s} > z_{1-\alpha}, \theta \neq 0]$. The decision maker seeking to maximize EG should use $z_{1-\alpha} = 0$ or $\alpha = 0.50$, i.e., should roll out any treatment with a positive observed effect regardless of significance level. This decision rule is consistent with prior analyses by Stoye (2009), Manski and Tetenov (2016), and Feit and Berman (2019).

Because $Pr(\theta \neq 0 \mid \widehat{\theta}/\widehat{s} > z_{1-\alpha}) = 1 - FDR$ from one-sided tests, Equation (12) implies

$$EG = Pr(\widehat{\theta}/\widehat{s} > z_{1-\alpha})\cdot[1 - FDR] \cdot E\left[\theta \mid \widehat{\theta}/\widehat{s} > z_{1-\alpha}, \theta \neq 0\right]. \tag{13}$$

Hence, the expected gain in effectiveness is the probability of declaring a discovery multiplied by the probability that a discovery is true rather than false multiplied by the expected true effect given that it is nonnull and its estimate is declared a discovery. Consequently, given a set of significant findings, the expected boost in effectiveness decreases with the FDR. We quantify how these expected gains are affected by π_0, μ, σ, $\widehat{s}^2$, and α in Section 7.

Of course, decision makers may not seek to simply maximize EG without taking into consideration the cost of deploying false discoveries. Switching from the baseline to the newly discovered treatment may trigger switching costs. For many experiments, the latter will be low, like changing the background color of a web page. However, for some, it may be quite substantial, like building and rolling out the infrastructure to enable a new shipping policy. Switching costs are incurred for both false and true discoveries, with expected frequency $\alpha\pi_0 + \Phi\left(\frac{\mu - z_{1-\alpha}\cdot\widehat{s}}{\sqrt{\widehat{s}^2+\sigma^2}}\right)(1-\pi_0)$. In addition, top management may worry that deploying false discoveries will harm their efforts to instill a test and learn culture, and the analytics team may worry about their credibility within the firm. These costs of disappointment are incurred only for false discoveries, with expected frequency $\alpha\pi_0$. Furthermore, decision makers and analysts may care more about avoiding losses than making equally sized gains, and hence, they may want to sharpen the significance level to avoid interventions with a positive observed effect but a true negative effect. In short, depending on their cost of switching and disappointment and their level of loss aversion, decision makers and analysts may want to use a value of $\alpha < 0.50$ in their decision rule.

Implementing a false discovery results in forgoing the expected gain from experimentation and incurring various costs. The consequences may vary across organizations. For instance, large organizations with high-volume traffic to their websites and plenty of analytics and engineering resources will be able to quickly detect the lack of improvement and will have the resources to test new ideas and implement improvements. Hence, they will forgo the gains of a true discovery for a shorter period compared with smaller organizations. However, the same organizations operating on a massive scale typically care about even very small forgone improvements.

4. Data

4.1. Research Setting

Our data come from Optimizely, an online A/B testing platform. It helps experimenters with designing, delivering, monitoring, and analyzing different versions of web pages. 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 web page, called web page variations. When an online user visits the experimenter's website, the platform assigns this visitor to one of the variations randomly, 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 monitored actions are chosen by the experimenter and are called "goals." These goals can include engagement, clicks, page views, revenue, or other actions defined by the experimenters. In this study, we focus on engagement as the goal, which is defined as clicking anywhere on the tested variation and is the default and most popular goal on the platform. This allows us to compare performance on the same goal across experiments and results in the largest set of experiments for us to study.

The platform logs the number of unique visitors and the number of unique engagement clicks, also called conversions. The conversion rate of each variation is defined as the number of conversions divided by the number of visitors. 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 with the baseline, and statistics are computed relative to the baseline.

The platform reports the result of a one-sided t test comparing each variation with the baseline. The tests are performed at 5% significance and called "chance to beat the baseline." Figure 2 presents the dashboard displayed to the experimenter. The test statistic is only displayed after the numbers of visitors to the baseline and to the variation both reach 100.

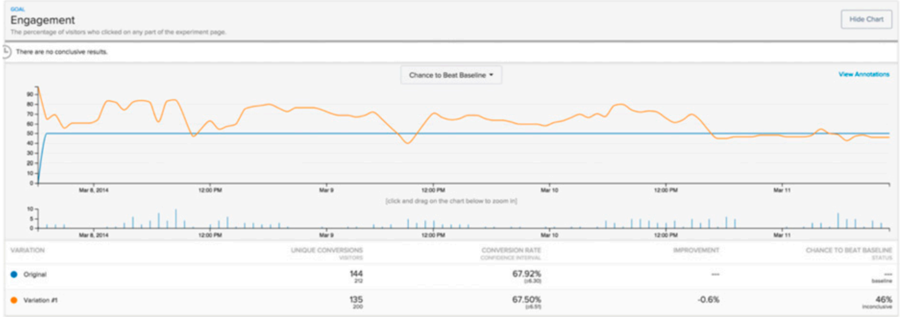
4.2. Set of Experiments Studied

Our raw data contain all 8,598 experiments that were registered on the platform during the month of April 2014. The data contain 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.

We exclude experiments that have one or more of the following characteristics.

1. Having all conversion rates at 100% or all conversion rates at 0%
2. Having a conversion rate above 100%
3. Having less than 100 visitors to the baseline or to all its variations
4. Not having engagement as a goal
5. Ending after November 30, 2014 (the end of our data window)
6. Having no traffic for six consecutive days or all traffic assigned to one variation for six consecutive days

Figure 2. (Color online) Experimenter Dashboard: Overview



The percentage of visitors who clicked on any part of the experiment page.

There are no conclusive results.

VARIATION	UNIQUE CONVERSIONS VISITORS	CONVERSION RATE CONFIDENCE INTERVAL	IMPROVEMENT	CHANCE TO BEAT BASELINE STATUS
● Original	144 212	67.92% (±6.30)	---	--- baseline
● Variation #1	135 200	67.50% (±6.51)	-0.6%	46% inconclusive

Table 4. Summary Statistics of Experiments

	Mean	Median	Standard deviation	Min	Max
No. Nonbaseline Variations	1.83	1	2.26	1	64
Total Goals	4.18	3	5.72	1	158
Past No. Exp.	248.84	97	390.50	1	2,917
Length	26.04	15	28.23	1	162

Notes. $N = 2{,}766$. Values are computed on the last day of the experiment.

The first three criteria remove experiments with poorly measured effect sizes or poor statistical inference. The final criterion excludes experiments that very likely were terminated de facto by reconfiguring the website before the experimenter notified the platform about the experiment's termination. Knowing when an experiment was ended is necessary for us to determine whether an experiment's result would have been declared statistically significant.

Our final data set consists of 4,964 effect sizes from 2,766 experiments run by 1,349 experimenter accounts.[6] Thirty-six percent of experiments have more than one nonbaseline variation. Only 15% of the experiments list engagement as the only goal, and it is possible that some experimenters pursued a primary goal other than engagement. We take the number of variations and number of goals into account in our analyses.

4.3. Descriptive Statistics

Table 4 reports several characteristics of the experiments. The median number of variations excluding the baseline was one, and the median number of goals was three. Experimenters varied quite a bit in the number of prior experiments they had run on the platform, with the median being 97. On the last day, the typical (median) experiment had run for 15 days.

Table 5 reports descriptive statistics for the 4,964 variations we are analyzing. The variables are sample size, effect size, lift, and z score.

Sample size includes the number of visitors to the baseline and nonbaseline variation. The effect size of a nonbaseline 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 2). The asymptotic z score is computed as the effect size divided by its standard error computed as $\sqrt{\frac{c_b(1-c_b)}{n_b}+\frac{c_v(1-c_v)}{n_v}}$, where c_b and c_v are the observed conversion rates for the baseline and the variation, respectively, and n_b and n_v are the sample sizes, respectively.

Figure 3 shows that effect sizes and especially z scores exhibit long tails. This is confirmed by their kurtosis being much higher than that of a normal distribution (30 and 451, respectively, versus 3) and by Shapiro–Francia tests rejecting the null that either variable is normally distributed ($p < 0.0001$). The red horizontal line in the histogram of p-values crosses the vertical axis at 1.8%. Because there are 40 bins, the mass under that line amounts to 72% of the p-values.

Reflecting the experience of many practitioners using A/B tests, the effect sizes tend to be very small and frequently nonsignificant; only 26% are significant at $\alpha = 10\%$, 20% are significant at $\alpha = 5\%$, and 13% are significant at $\alpha = 1\%$.

4.4. Comparison with More Recent Experiments

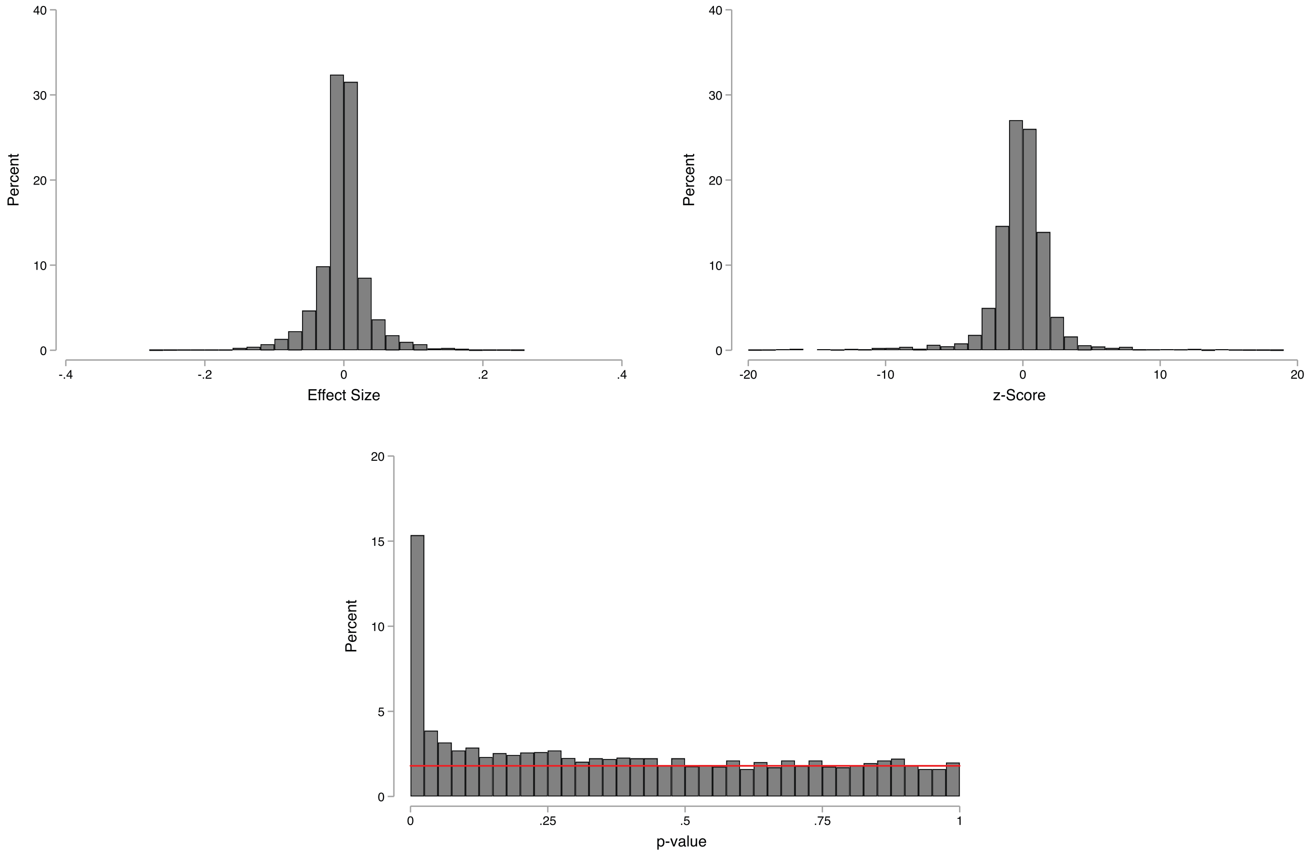
The experiments from 2014 that we analyze are similar to those run in 2017 and 2018 on the same platform, as reflected in basic descriptive statistics for 21,836 experiments run between November 2016 and September 2018 (Thomke 2020, pp. 110–112). The fraction of experiments with only a single nonbaseline variation barely changed (64% versus "about 70%"), as did the mean number of nonbaseline variations (1.8 versus 1.5). The average run time of experiments increased slightly from 3.7 to 4.4 weeks. The four industries that experimented the most remained the same: retail, media, hi-tech, and professional or financial services. Thomke (2020, p. 111) reports that "19.6% of all experiments achieved statistical significance on their primary metric." We find that 26% of all experiments in our data reached 5% significance (one sided) on engagement.[7] As in the 2016–2018 data, this number splits

Table 5. Summary Statistics of Effects

	Mean	Median	Standard deviation	Min	Max
Sample Size (base + focal)	61,009	3,781	556,159	201	34,724,196
Effect Size	−0.001	−0.001	0.043	−0.639	0.464
Lift	0.023	−0.001	0.749	−0.821	24.374
z score	0.105	−0.067	6.939	−80.432	196.084

Notes. $N = 4{,}964$. Values are computed on the last day of the experiment.

Figure 3. (Color online) Histograms of Effect Sizes, *z* Scores, and *p*-Values



Note. $N = 4{,}920$, covering 99.1% of the data.

almost evenly between positive and negative effects. In short, even though the performance metric on which the tests are performed differs, the pattern of significance did not change markedly.

5. Results

5.1. How Many Effects Are True Nulls?

Table 6 reports the estimates of π_0 from Methods A, B, and C. For A and B, it also reports estimates of the other model parameters. Column B/C reports estimates of Model B where π_0 is restricted to the estimate obtained from Method C.

Method A estimates π_0 to be 67% (95% C.I., 64%–69%). As noted earlier, this method does not take into account how standard errors differ across tests. Method B does and estimates π_0 to be 80% (95% C.I., 78%–81%). Method B has less restrictive assumptions and fits the data markedly better ($\Delta - 2LL = 2{,}230$).

Whereas Methods A and B assume normality for the nonnull effects, Method C does not make any parametric assumptions on those effects. Its estimate of π_0 is 72% (95% C.I., 65%–78%). Given the differences between the estimates from Methods B and C, we reestimate the model in Method B after restricting π_0 to the value estimated from Method C, corresponding to the red horizontal line in the histogram of *p*-values in Figure 3. This restricted model fits worse ($\Delta - 2LL = -73$) but produces very similar estimates of μ and σ. Hence, we believe that the 72% estimate of π_0 is credible across methods.

True nulls amounting to 70% of all effects may sound high, yet it is consistent with an earlier report that the true null rate in experiments conducted on Microsoft's search engine Bing was over 80% (Deng 2015). Also, it compares favorably with academic

Table 6. Estimates of π_0 and Other Model Parameters

	Method			
	A	B	C	B/C
π_0	0.6654*** (0.0145)	0.7976*** (0.0087)	0.7162*** (0.0349)	0.7162
μ	−0.0036* (0.0018)	−0.0032 (0.0026)		−0.0030 (0.0022)
σ	0.0714*** (0.0017)	0.0759*** (0.0022)		0.0693*** (0.0017)
s	0.0133*** (0.0004)			
$-2LL$	−20,145	−22,375		−22,302

Note. $N = 4{,}964$; s.e. values are in parentheses.
*$p < 0.05$; ***$p < 0.001$.

psychology where the true null rate has been estimated to be about 90% (Johnson et al. 2017b).

Both Methods A and B allow the first component of the mixture to capture the spike at zero. Rescaling effect sizes by their standard error results in a density of z scores with longer tails compared with effect sizes. It is these longer tails that result in a higher estimate of σ and π_0 in Method B compared with Method A.

Method C is less sensitive than Method B to extreme z scores because their transformation into p-values using the normal cdf squeezes extreme z scores into the unit interval. Consequently, Method C does not require as large a π_0 to account for extreme z scores as Method B does.

5.2. How Many Discoveries Are False?

Table 7 reports the FDRs for two-sided tests at five levels of significance. Method B implies an FDR of 37% at $\alpha = 10\%$, which was the default level used by the platform to declare significance. At 5% significance, the FDR is 25%, meaning that as many as one of four significant results remains a false discovery. The FDR decreases as one tightens the significance level further but remains higher than α.

Remember that Method C produces a lower estimate of π_0 than Method B does (72% instead of 80%). This results in a lower estimate of the FDR at each significance level α as implied by Equation (1) and Figure 1. Using the same estimate of π_0 from Method C to estimate the FDR using the model of Method B leads to nearly identical FDR estimates in Method B/C. The FDR equals 28% at $\alpha = 10\%$ and 18% at $\alpha = 5\%$.

An FDR of 18% for website A/B tests conducted at $\alpha = 5\%$ may seem surprisingly high. Yet, it compares favorably with FDRs for tests at the same level of significance in medical research, which experts believe range between 20% and 50% (Benjamini and Hechtlinger 2013), and with FDRs in psychology, where analyses of three different bodies of test results reported FDRs of 41%, 58%, and 81% (Gronau et al. 2017).

5.3. Is Low Power the Culprit for the High FDRs?

Variation in the FDR is induced not only by differences in α and π_0 but also by differences in the power $1-\beta$. We therefore investigate how the FDR varies with power, i.e., the probability of correctly rejecting the null hypothesis at a specific α.

Table 7. FDR (Percentage) at Various Significance Levels

	Method		
α (%)	B	C	B/C
20.0	51.0	40.0	40.7
10.0	36.7	27.8	28.0
5.0	24.6	18.0	18.3
1.0	8.2	5.5	6.0
0.1	1.5	0.8	1.2

Table 8. Power (Percentage) at Various Significance Levels

	Method		
α (%)	B	C	B/C
20.0	76.9	77.1	74.9
10.0	69.0	66.8	66.1
5.0	61.3	58.6	57.4
1.0	44.8	44.2	40.3
0.1	26.3	31.9	21.2

Using the estimates from Methods B, C, and B/C, and rearranging Equation (1), we compute the average power as

$$\widehat{Power}(\alpha) = \alpha \frac{\widehat{\pi}_0}{1-\widehat{\pi}_0} \frac{1-\widehat{FDR}(\alpha)}{\widehat{FDR}(\alpha)}. \tag{14}$$

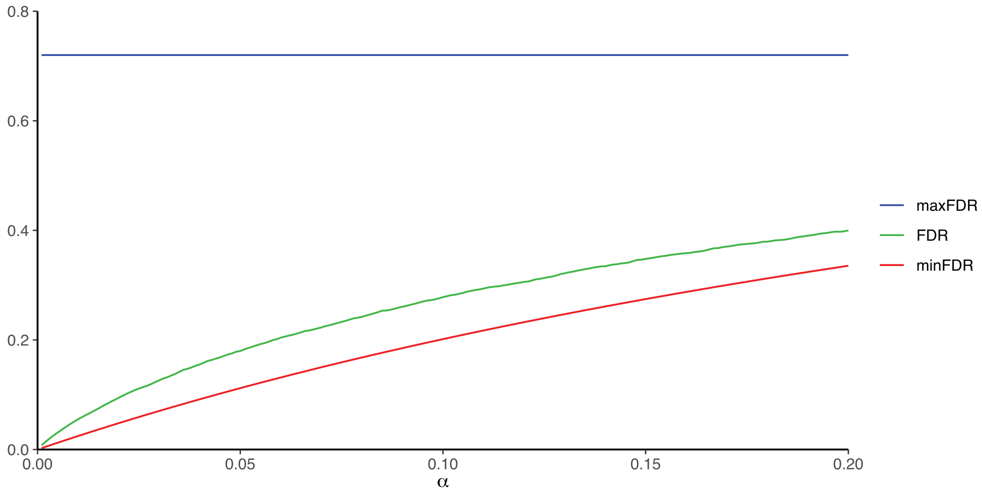
The results are reported in Table 8. The power of the tests at 20% or 10% significance is 66% or higher. Hence, the high FDRs of these tests do not result from low power. Conversely, the power at 1% or 0.1% significance is only 21%–45%, yet the FDRs are only 1%–8%. Hence, power does not seem to be the main determinant of the FDR among the test results. Rather, it is the high value of π_0.[8]

Figure 4 provides additional evidence that lack of power is not a major contributor to the FDR in our data. The middle line shows how the FDR computed using Method C varies with α. This line reflects the FDR at the actual power levels, which of course, also vary with α. The bottom line (minFDR) shows how the FDR varies in a population of experiments where power is 100% using Equation (1). The top line (maxFDR) shows how the FDR varies in a population of experiments where the power is at its minimum, which is α in unbiased tests, resulting in an FDR of π_0. The fairly narrow gap between the middle and bottom lines again indicates that lack of power is not the main driver behind the FDR values we observe. Even if power was 100%, the FDR would still be 11% at $\alpha = 0.05$ and 20% at $\alpha = 0.10$.

5.4. Heterogeneity in the FDR

The results so far pertain to the average experiment. Conceivably, the fraction of true nulls (π_0), the distribution of the nonnulls (μ and σ), and the precision of the tests ($\widehat{s}_i$) may vary across experiments. This section, therefore, explores how the FDR and its key drivers vary with four traits of experiments: (i) the past experience of the experimenter, (ii) the number of goals in the experiment, (iii) the number of variations in the experiment, and (iv) the industry of the experimenter. We do not consider sample size because its effect on the FDR is already accounted for through $\widehat{s}_i^2$.

Figure 4. (Color online) How FDR at Actual, Minimum, and Maximum Power Varies with α When $\pi_0 = 72\%$



The experience of the experimenter may be associated with the FDR in various ways. Greater experience in running experiments may help the experimenter to generate more ideas that are not null (decrease π_0) and nonnulls that have bigger effects (μ farther away from zero or higher σ). However, causality may also run the other way; people who had more positive experiences when implementing A/B tests (lower FDR) may run more experiments. This would also result in experience being associated with a lower π_0, μ being farther away from zero, a higher σ, and a lower FDR. Finally, the association between experience and FDR could be negative if experienced experimenters run out of ideas and start "scraping the bottom of the barrel."

Our analysis focuses only on effects in terms of engagement. However, 85% of the experiments tracked more than one goal. It is fair to assume that the greater the number of goals tracked, the less likely it is that boosting engagement is the main objective of the intervention tested. Additionally, because an intervention meant to affect a goal other than engagement is more likely to have a true null effect in engagement, a larger number of goals may be associated with higher π_0 and hence, a higher FDR.

The third variable we investigate is the number of variations in the experiment. Experimenters testing many variations concurrently may be "scraping the bottom of the barrel," resulting in a higher π_0, a lower average effect among nonnulls (lower μ), and fewer ideas that really move the needle (lower σ). Conversely, such experimenters may be "swinging for the fences" and be "going for the long tail" (Azevedo et al. 2020), which could translate into a large fraction of nulls (higher π_0) but also more outliers among the nonnulls (higher σ). Note that including a greater number of variations may result in less traffic per variation, which would harm the power and increase the FDR, but that mechanism is already accounted for through $\widehat{s_i^2}$.

We extend the model in Method B and make π_0, μ, and σ a function of the three covariates. We use a logit transformation for π_0 because it is bounded between zero and one, we use a log transformation for σ because it is nonnegative, and we use a linear expression for μ as it is unbounded. We measure the experimenters' experience as the number of experiments they ran previously on the platform. We use a log transformation, which is consistent with the learning curve effect and protects the results from being affected by outliers in the highly skewed distribution (see Table 4). Similarly, we use a log transformation on the number of goals to protect the results from artifacts because of outliers. Finally, we mean center the three covariates so the intercepts map roughly into the estimates from Method B in Table 6. Five experiments accounting for nine observations are lost because of missing values for the number of past experiments.

Table 9 shows the results. Adding the nine parameters notably improves the fit compared with Model B

Table 9. Method B Mixture Model with Parameters as a Function of Covariates

	logit(π_0)	μ	log(σ)
Intercept	1.3202***	−0.0020	−2.6560***
	(0.0572)	(0.0027)	(0.0312)
log(*Past No. Exp*)	−0.1136**	0.0059***	0.0045
	(0.0367)	(0.0015)	(0.0215)
log(*Total Goals*)	0.2251**	0.0110**	0.1614***
	(0.0714)	(0.0035)	(0.0419)
No. Nonbaseline Variations	−0.0517***	−0.0010***	−0.0338***
	(0.0066)	(0.0001)	(0.0058)

Notes. N = 4,955 from 2,761 experiments. $-2LL = -22{,}516$; s.e. values are in parentheses.
$^{**}p < 0.01$; $^{***}p < 0.001$.

reported in Table 6 ($\Delta - 2LL = 182, df = 9, p < 0.001$; ΔBIC = 117). Greater past experience is associated with a lower π_0 and an above-average μ, a combination expected to translate into a lower FDR. In contrast, a larger number of goals is associated with not only a higher π_0 but also a higher μ and σ. Whereas a larger fraction of true nulls is associated with a higher FDR, a higher mean and a higher variation among the nonnulls are associated with greater power and hence a lower FDR. Overall, this combination does not translate into an unambiguously upward or downward shift in the FDR. Finally, a larger number of nonbaseline variations is associated with a lower π_0, a lower μ, and a lower σ, which again do not translate into a clear upward or downward shift in the FDR. Adding fixed effects for the four main industries (retail, media, hi-tech, and financial and professional services) and adding an account-specific random effect on π_0 barely affect the point estimates, but the coefficient of the number of goals on π_0 loses statistical significance because of a higher standard error (see Online Appendix B.1).

To gain clarity in how the three covariates are associated with the FDR, rather than just π_0, μ, and σ, we compute the FDR by tercile of each covariate. Results are reported in Online Appendix B.2. The main insight is that the FDR decreases monotonically as we move from the bottom to the top tercile in experience but does not show a monotonic change across terciles in the other covariates. For instance, the FDR at 5% significance is 24% in the first tercile of experience (1–40 experiments), decreases to 19% in the middle tercile (41–202 experiments), and reaches 13% in the top tercile (203 or more experiments). For the other two covariates, the FDRs in the bottom and top terciles vary only between 16% and 17%.

Applying Method C for the data from each of the top four industries separately shows that π_0 and the FDR values vary little across these industry verticals. However, they have a lower π_0 and lower FDRs than the remaining industries. Reasons could be that these four industries have greater experience and greater sample size than average (see Table B.2 in the online appendix).

5.5. The Probability That a Particular Effect Is Null

Managers may want to know the probability that a particular effect is a true null, taking into account its standard error but regardless of a specific significance cutoff. This is known as the local fdr and is given by Equation (3), which we present again with subscripts for observation *i*:

$$Pr(\theta_i = 0 \mid z_i) = \frac{\pi_0 f_0(z_i)}{f(z_i)}. \qquad (15)$$

We compute the local fdr for each observation using the estimates of π_0 and $f(z_i)$ from Method C.

Figure 5. (Color online) Probability That an Effect Is a True Null (Local fdr) by *z* Score

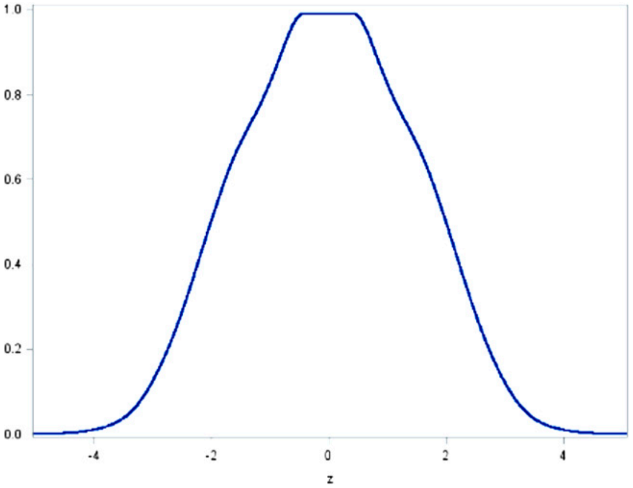


Figure 5 plots the local fdr against the *z* scores between −5 and 5. The left panel of Figure 6 plots the local fdr against the observed effect sizes between −0.3 and 0.3, whereas the right panel shows the boxplot by 5% quantiles. The left panel shows dispersed dots rather than a single smooth line because of differences in precision across effect sizes. These plots give analysts and managers a sense of how likely a particular effect is a true null. Overall, of all observed effects, only 19.2% have a local fdr of 50% or less, and only 10.1% have a local fdr of 10% or less.

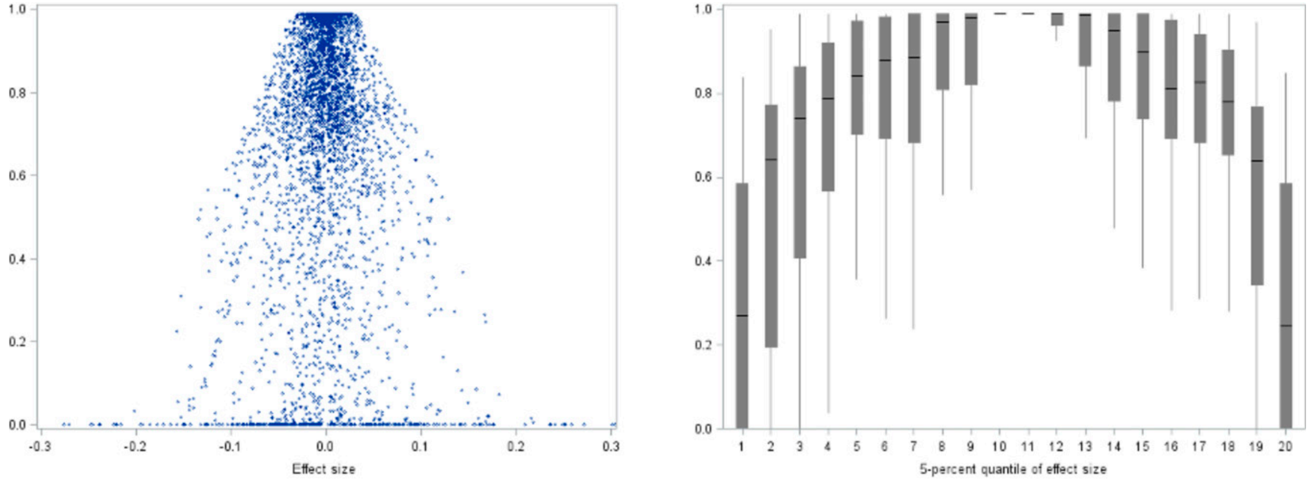
6. Robustness to Key Assumptions

In this section, we investigate features of the data or behaviors of the experimenters that might inflate or deflate the estimates of π_0 and the FDR. The first possible concern pertains to the fact that we assess engagement effects, whether or not that was the primary goal of the experiment. The second and third concerns pertain to possible malpractice by the experimenters in declaring test results to be discoveries: improperly conducting multiple comparisons and improperly testing hypotheses after peeking at the data (Benjamini and Hochberg 1995, Johari et al. 2017). Finally, we also report the FDRs if analysts used one-sided tests on positive effects only rather than two-sided tests to declare discoveries.

6.1. Experiments with Engagement as the Only Goal

Our analyses so far assumed that experiments were designed to test treatment effects in terms of engagement. If many experiments were designed to affect a goal other than engagement, then our estimates of π_0

Figure 6. (Color online) Probability That an Effect Is a True Null (Local fdr) by Effect Size



Notes. (Left panel) Observed effect size. (Right panel) Five-percent quantiles of effect sizes. Boxes cover the interquartile range of local fdrs in a given quantile of effect size.

and the FDR values would be overly pessimistic. Say many interventions were designed to boost sales and not engagement. In that case, one would expect less systematic difference in engagement across arms of the experiment than when the goal was engagement, which would translate into a higher π_0 and a higher FDR assuming no change in power. We therefore repeat the main analyses but now restricted to the 682 (13.8%) effect sizes of the 424 (15.3%) experiments tracking only engagement. Online Appendix B.3 reports the equivalent of Tables 6 and 7 for this subset of effects. The parameter estimates from Method A are similar to those for the full data set. Notably, $\widehat{\pi}_0$ is only slightly lower (63% versus 67%) and well within the error margin. The parameter estimates from Method B show greater differences. Notably, $\widehat{\pi}_0$ is somewhat lower (71% versus 80%), and this corresponds to somewhat lower FDRs for tests at the 20%, 10%, and 5% significance levels. The FDRs for tests conducted at 1% or 0.1% significance, in contrast, are not affected. Similarly, Methods C and B/C result in only slightly lower π_0 (69% versus 72%) and FDR values.

In short, the engagement-only experiments exhibit only moderately fewer true nulls and false discoveries, with the latter difference vanishing at more stringent significance levels. The main conclusions implied by the analyses in Sections 5.1 and 5.2 hold.

6.2. Click Instead of Engagement

We repeat the analysis for the 1,065 experiments where a click on a specific link in the page was designated as the goal, which yields a total of 1,985 effects.

The mean and median effect sizes are both virtually zero. An FDR analysis using Method C produces an estimate of π_0 of 75%, which is similar to that for engagement. The FDR estimates are very similar as well (36%, 23%, 15%, 4%, and 1%).

6.3. Improper Multiple Comparisons

When experimenters test more than one variation in an experiment, they may declare the result of the experiment to be a discovery if any one of the variations yields a significant result based on the p-value unadjusted for multiple comparisons. This behavior inflates the effective type I error rate above its nominal level (e.g., 0.05), which in turn, inflates the FDR. Specifically, if k effects are tested with the same levels of α and β, then the type I error rate in declaring at least one effect significant when all are actually null equals $1-(1-\alpha)^k$. This is traditionally referred to as the family-wise error rate. Also, the type II error rate equals β^k. Consequently, the FDR equals (Ioannidis 2005, Maniadis et al. 2014)

$$FDR(k) = \frac{\pi_0\left(1-(1-\alpha)^k\right)}{\pi_0\left(1-(1-\alpha)^k\right)+(1-\pi_0)(1-\beta^k)}. \tag{16}$$

Taking derivatives shows that $FDR(k)$ increases in k when $1-\beta > \alpha$. Unless an experiment is massively underpowered for the chosen significance level, this condition will be met.

The FDR values reported were computed for the case where experimenters properly declared discoveries variation by variation, an assumption that need not hold. Hence, we investigate the following question. What would the FDR be in our set of experiments if experimenters engaged in improper multiple comparison testing and declared an experiment as significant if any variation reached significance? To answer this question, we use the estimates of π_0 and β

Table 10. The Increase in FDR Because of Improper Multiple Comparisons

		FDR(*k*) (%)					
α(%)	Power	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
20.0	0.771	40	49	56	60	63	65
10.0	0.668	28	35	42	47	51	55
5.0	0.586	18	23	28	33	37	41
1.0	0.442	5	7	8	10	12	13
0.1	0.319	1	1	1	1	1	2

(i.e., 1 − power) from Method C reported in Table 8 and enter those in Equation (16). The values reported in Table 10 show how much improper multiple comparisons would inflate the FDR. For significance levels between 1% and 10%, the FDR roughly doubles when multiple comparisons involve five or more variations.

This analysis relies on estimates of average power to obtain β values. Online Appendix B.4 describes an alternative approach, which produces very similar results without relying on average power.

The presence of multiple variations can lure experimenters into conducting multiple comparisons, and doing so improperly is associated with higher FDRs. This may seem at odds with the finding in Section 5.4 that the FDR did not increase systematically with the number of variations. This apparent paradox is explained by the fact that the FDR values reported in Section 5 assume that experimenters do not engage in *improper* multiple comparisons (i.e., they do not declare the result of the experiment as a discovery if any of its variations yield a significant result but rather, declare discoveries variation by variation). Hence, FDR values reported in Section 5 need no correction for multiple comparisons.

6.4. Optional Stopping

A second possible malpractice that can make FDR estimates overly conservative (i.e., optimistic) is that experimenters peeked at the results and stopped the experiment only after it reached a desired significance level. This behavior is often referred to as “data peeking” or “optional stopping,” and it is a concern in the A/B testing world (Wald 1945, Pocock 1977, Johari et al. 2017).

Engaging in optional stopping results in declaring more findings significant than is justified. In terms of Table 3, a fraction u of the $m - S$ experiments that would be called insignificant without optional stopping now shifts to the left column. Following Ioannidis (2005), we assume that this fraction u is the same across rows, i.e., that it does not depend on whether the true effect is null or not. Optional stopping then changes the $FDR = E\left[\frac{F}{S}\right]$ to $FDR(u) = \mathbb{E}\left[\frac{F+u(m_0-F)}{S+u(m-S)}\right]$. Also, the expression in Equation (1) changes to (Ioannidis 2005, Maniadis et al. 2014)

$$FDR(u) = \frac{\pi_0(\alpha + u(1-\alpha))}{\pi_0[\alpha + u(1-\alpha)] + (1-\pi_0)(1-\beta+u\beta)}. \quad (17)$$

The u bias inflates the true type I error rate above its nominal level α to $\alpha + u(1-\alpha)$ but also increases the power. Taking the derivative of $FDR(u)$ shows that the net effect of the u bias on the FDR is positive when $\alpha < 1-\beta$ or equivalently, $FDR < \pi_0$. As noted, this condition likely holds in the great majority of cases.

Because the FDR values we reported use the nominal value of α, it is possible that these values are biased downward. To address this possibility, we assume that the FDRs on the penultimate day of the experiment are free of any such bias. If experimenters delayed stopping the experiment by one day in the hope of getting more statistical significance, either through a bigger effect size or a larger sample size, then the value of the last day may be biased but not that of the day before.

First, we express $FDR(u)$ as

$$FDR(u) = \mathbb{E}\left[\frac{F_{T-1} + u(m_0 - F_{T-1})}{S_{T-1} + u(m - S_{T-1})}\right] \approx \frac{\mathbb{E}[F_{T-1} + u(m_0 - F_{T-1})]}{\mathbb{E}[S_{T-1} + u(m - S_{T-1})]}, \quad (18)$$

where $T-1$ denotes the penultimate day and T is the final day, and the approximation follows Storey and Tibshirani (2003). Next, we use the property that $m_0 - F_{T-1}$ converges to $\pi_{0,(T-1)}m(1-\alpha)$ as m grows large. We then apply Method C to the data from the penultimate day to estimate $\pi_{0,(T-1)}$ and estimate u as $\frac{S_T - S_{T-1}}{m - S_{T-1}}$ (i.e., the fraction of effects declared not significant in the penultimate day that is declared significant in the final day). We then estimate $FDR(u)$ as

$$\widehat{FDR}(u) = \frac{\widehat{\pi}_{0,(T-1)}m(\alpha + \widehat{u}(1-\alpha))}{S_T}. \quad (19)$$

Table 11 reports three values for each nominal α level: (i) the u bias-inflated α, (ii) the standard FDR computed using Method C ignoring u bias, and (iii) $FDR(u)$, which takes the bias into account. The values are computed on the 4,672 effects from experiments

Table 11. The Increase in α and the FDR Because of Optional Stopping Delay of One Day

α(%)	$\alpha + u(1-\alpha)$ (%)	*FDR* (%)	*FDR*(*u*) (%)
20.0	21.1	39.7	42.9
10.0	11.0	27.4	31.0
5.0	6.1	17.5	21.8
1.0	1.4	5.3	7.6
0.1	0.5	0.7	4.0

Note. $N = 4{,}672$.

running at least two days and having a sample size of 100 or more in both the baseline and the other focal variation on the penultimate day.

As expected, $FDR(u)$ is higher than FDR. However, the increase is small. This implies that the conclusions from the main analysis remain qualitatively valid even if all experimenters engaged in optional stopping by delaying the end of their experiments by one day.[9]

6.5. FDR from One-Sided Tests

Finally, we also compute the FDR for the case where analysts and decision makers declare discoveries only for positive effects and do so using one-sided rather than two-sided tests. Because Method C cannot be used for one-sided tests, we calculate the FDR using Method B only. The FDR is 52% at α = 10%, 38% at α = 5%, 14% at α = 1%, and 3% at α = 0.1%. As expected given that the estimate of μ is essentially zero, the FDRs from one-sided tests at significance level α are nearly identical to those from two-sided tests at $2 \cdot \alpha$ calculated using the same Method B (compare Table 7). Calculating the FDR using the nonparametric method of Pounds and Cheng (2006) produces similar values: 47%, 34%, 11%, and 2%.

7. How Can Firms Lower Their FDR?

Firms can use five main levers to lower the FDR: (i) reducing α, (ii) boosting the power by increasing the precision $1/\widehat{s}_i$, and identifying candidate interventions that are less likely to be true nulls and that are larger given that they are not null, which implies (iii) decreasing π_0, (iv) shifting μ away from zero, or (v) increasing the variance of true effects σ^2. Note that shifting μ and σ away from zero increases the odds of having a large true effect and hence also boosts the power of the test.

We start our analysis with quantifying the marginal effects of these parameters on the FDR and the expected gain from experimentation EG using one-sided tests. We do so using the estimates from Method B for π_0, σ, and μ; the observed $\widehat{s}_i$; and $\alpha = 0.05$. Calculating Equation (11) for each observation and then integrating over the empirical distribution of $\widehat{s}_i$, we obtain FDR = 38% and EG = 0.0050 or a lift of 1.10%, which places EG at the 70th percentile of all observed lifts and the 37th percentile of all positive observed lifts. Using these empirically informed values of the FDR and EG as baseline, we then vary each parameter one by one. Table 12 shows that π_0 is the dominant contributor to improvements in FDR and EG. It also shows that, except for changes induced by tightening α, lowering the FDR is associated with increasing the EG.

Table 12. Marginal Effects of Key Parameters on the FDR and the Expected Gains from One-Sided Tests

Parameter	Change (%)	Change FDR (%)	Change EG (%)
π_0	−10	−24.9	39.4
σ	+10	−2.5	12.7
μ	+10	−0.3	0.6
$\hat{s}_i$	−10	−2.4	2.1
α	−10	−5.4	−0.6

Changing π_0, μ, and σ requires changing the quality of the interventions being generated. That may be difficult to achieve directly. We therefore discuss seven possible strategies to change the design of the A/B test and lower the FDR. The first three involve only α and the sample size, and our empirical estimates imply that they do not hold much promise. The next three rely on using two-step experimental designs or two-step analyses of multiple variations to identify the best one. The last approach also involves picking the best out of several variations, but the assessment occurs in a single step and does not require special designs or holdout samples.

7.1. Tightening Significance Level α or Boosting Power Through Sample Size

Because the significance level α is fully within the experimenter's control, one might conclude from Equation (1) that the simplest way of reducing the FDR is to lower α (Benjamin et al. 2018). Indeed, our analyses indicate that the FDR decreases from 18% to 1% when one sharpens the significance level from $\alpha = 0.05$ to $\alpha = 0.001$. However, decreasing the FDR by sharpening α comes at a cost because the type I and type II error rates are not independent. Specifically, using the average values reported in Tables 7 and 8, it is far from obvious that a 17-point decrease in FDR from 18% to 1% is worth a 27-point decrease in power (i.e., the ability to correctly detect true nonnulls) from 59% to merely 32%.

Another way to reduce the FDR is to boost the power of experiments (e.g., Camerer et al. 2018). Doing so for a given α and true effect size requires a larger sample size and hence likely a longer run time. Power can also be increased by reducing the error variance through stratification. This requires not only more sophisticated algorithms but often also relevant covariates (Berman and Feit 2019, Bhat et al. 2020). We do not expect either procedure to markedly improve the FDR in the experiments we analyzed because they already had nearly adequate power: roughly 60%–70% at both $\alpha = 0.1$ and $\alpha = 0.05$. Moreover, our estimate of $\pi_0 = 0.72$ implies that even if power was increased to 100%, the FDR would not fall below 20% at $\alpha = 0.1$ and 11% at $\alpha = 0.05$.

Finally, experimenters may target a specific level of FDR when planning their experiments. Because the goal of A/B testing is to identify better practices, experiments need to be designed to avoid type II errors. In contrast, for diagnostic testing and decision making, avoiding type I errors may matter only to the extent that it avoids making false as opposed to true discoveries. If or when the error rate α is less important than the FDR, why not design experiments accordingly?

Experimenters would, in consultation with decision makers or after a formal analysis of the value of information, choose target levels of power and FDR and then identify the level of α implied:

$$\alpha = Power \frac{FDR}{1 - FDR} \frac{1 - \pi_0}{\pi_0}. \tag{20}$$

Table 13 shows the significance levels required for three levels of power ranging from 70% to 90% and for four levels of FDR ranging from 5% to 30%, assuming $\pi_0 = 72\%$. Note that the significance level α does not need to be sharper than 1%.

Given α and β, the next step is to compute the required sample size as usual. Calculations reported in Online Appendix B.5 show that if the true effect is at the 90th percentile of the observed effect sizes in our data, then the current median sample size per variation (roughly 1,900) is almost sufficient for a test with $\alpha = 0.10$ and power = 80%, which results in an FDR of 24% when $\pi_0 = 72\%$. However, if the true effect is only at the 75th percentile, then the required sample size is 10 times larger than the current median. If the effect is just better than two-thirds of all observed effects, then the sample size must be 50 times the median.

In short, except for large effects, the three approaches relying only on changing α or sample size are not very promising venues to lowering the FDR. Smarter approaches are desired, even for firms enjoying massive traffic but chasing very small effects.

7.2. Two-Step Testing

Replication is a straightforward way to reduce the number of false discoveries. Instead of declaring a discovery based on a single experiment, one can (1) take significant results as "potential discoveries," (2) run a replication of the first experiment, and (3) declare the effect a discovery only if it is again statistically significant in the same direction. There are at least three methods to operationalize this idea.

Table 13. Levels of α Required for Combinations of Power and FDR When $\pi_0 = 72\%$

	FDR			
Power (%)	0.05	0.10	0.20	0.30
90	0.018	0.039	0.088	0.150
80	0.016	0.035	0.078	0.133
70	0.014	0.030	0.068	0.117

The simplest way is to plan a replication only after a significant effect has been observed. This decreases the chance of a false positive from α to α^2. The downside is that without increasing the sample size, the power decreases from $1 - \beta$ to $(1 - \beta)^2$. It is, therefore, preferable to combine the two z values into a single test, which amounts to reducing the expected standard error by a factor of $\sqrt{2}$.

If one is precommitted to two-step testing, then further efficiency or power gains can be made by (1) running multiple candidate interventions in a first "screening" or "pilot" stage, (2) replicating only the largest or the most significant effect in a second "validation" stage, and (3) optimally splitting the total sample across the first- and second-stage experiments. One can also set different levels of α and β in each stage so as to achieve a given FDR at minimum sample size. This two-stage approach has attracted considerable attention (Zehetmayer et al. 2005, Zehetmayer and Posch 2012, Sarkar et al. 2013, Deng et al. 2014). Two main insights are that two-stage designs give a big boost in power and reduction in FDR over single-stage designs and that splitting subjects equally across stages is not quite optimal.

A variant is to run only a single experiment with multiple interventions, split the sample into an exploratory part and a validation part, use the exploratory subsample to identify the most promising candidate intervention, and then use the other subsample to validate and assess that candidate (Anderson and Magruder 2017). This variant of the two-stage approach can have the benefit of a compressed time window but only if having enough subjects for both subsamples does not require running the test just as long as in a genuine two-stage design.

The key benefit of all the two-step approaches involving multiple treatments is that they allow one to identify the most promising out of several candidates in terms of effect size. This in turn increases the power and reduces the FDR because the largest observed effect is less likely to come from the null and is more likely to reflect a large true effect. Consequently, this approach is especially valuable when the distribution of true effects exhibits long tails (Azevedo et al. 2020). As the histogram of observed effects and formal tests of kurtosis suggest and as the estimates of π_0 and σ indicate, the true effects in our data indeed exhibit long tails.

7.3. Pick the Best Without Testing

As our discussion of the expected gains from experimentation noted, these will be maximized by simply declaring any positive effect a discovery. This amounts to using α = 50% and will of course come with a high FDR. In a simple A/B test, the FDR will equal the fraction of true nulls. However, just as with a preplanned two-step study, an A/B test in which one picks the best out of multiple treatments will result not only in a higher EG but also a lower FDR. We therefore investigate "pick the best without any testing" as another possible improvement over traditional A/B testing.

7.4. Simulation Analysis

In this section, we use simulation to compare the performance of the replication approach against the traditional one-shot approach and pick the best without testing. We do so on two criteria: the FDR and the expected gain from experimentation EG. We expect replication to dominate the traditional A/B test on both metrics. Compared with pick the best, it is obvious that replication does worse on EG and better on FDR. Hence, the decision to use replication versus pick the best will depend on the relative gap in performance on each metric and how much the decision maker wants to avoid false discoveries even if this comes with forgoing gains.

We proceed as follows. We simulate data for 5,000 experiments, where in each experiment, we fix the conversion rate of the baseline and draw conversion rates for $k = 1, \ldots, 10$ variations by adding effect sizes, which are drawn according to a mixture model where the effect is zero with probability π_0 and drawn from $\mathcal{N}(\mu, \sigma^2)$ with probability $1 - \pi_0$. The sample size for each experiment is drawn from a log-normal distribution with mean and variance based on the empirical distribution of the 2,766 experiments we study.

Given the parameters of each experiment (sample size, and baseline and variation conversion rates), we simulate Bernoulli draws as the conversion of each member of the sample. We then calculate the observed effect size and pick one of the variations as the one being implemented in the field based on the three approaches we compare. For the traditional approach, we pick any variation that passes a one-sided significance test with $\alpha = 0.05$ for implementation. For the replication approach, we pick the variation with the largest observed effect size; test whether it passes a one-sided significance test with $\alpha = 0.05$; if it does, replicate the baseline and the variation Bernoulli draws; and measure the effect in the replication. The result is selected for implementation if the joint z score (using Stouffer's method) of the initial test and the replication passes the one-sided hypothesis test. Finally, for pick the best, we select the variation with the largest observed effect size and if it is positive, choose it for implementation. If no variation is selected for implementation, we assume that the experimenter stays with the baseline.

The sample size for each variation is set as follows. Given the sample size n of the experiment drawn as described and given the number of variations k in the experiment, the sample size of each variation is $\frac{n}{k+1}$ in the case of the traditional and pick the best approaches and $\frac{n}{k+3}$ for replication because an additional baseline and variation condition must be allowed for replication.[10]

Because we know for each effect whether it comes from the null or not, we calculate the FDR as the proportion of all implemented variations whose true effect is null. The expected gain EG is computed as the average implemented effect size across all experiments, i.e., zero if the baseline was selected or the true effect size of the variation if it was selected.

We set $\mu = -0.0032$ and vary $\pi_0 = (0.6, 0.7, 0.8)$ and $\sigma = (0.04, 0.08, 0.16)$. We present the results for the parameter values $(\mu, \pi_0, \sigma) = (-0.0032, 0.7, 0.08)$, which closely match our empirical estimates. Online Appendix C presents the results for the remaining combinations of parameter values.

Figure 7 shows the FDR for each approach with k ranging from 1 to 10. The replication approach has the lowest FDR, and it remains constant around 12%. The FDR for the traditional approach is about twice as high and increases from 24% for k = 1 to 29% for k = 10. The reason is that the sample size, and hence power, decrease with k. This also happens for the replication, but that gets neutralized by selecting the largest variation out of k. Clearly, pick the best, which is equivalent to a one-sided test with $\alpha = 0.50$, performs the worst. The FDR improves from $\pi_0 = 70\%$ when k = 1 to about 40% when k = 10. In short, replication dominates both alternatives in terms of FDR.

Figure 8 shows the expected gains EG for each approach with k ranging from 1 to 10. Consistent with Section 3, pick the best dominates. However, the gap with replication is rather small and remains fairly constant with k. Both dramatically outperform the traditional approach after k is larger than one. The reason is that both pick the best and replication benefit from selection when k increases.

In short, A/B tests with replication clearly dominate the traditional approach. Whether it dominates the pick the best approach depends on the trade-off that a decision maker makes between forgoing expected gains and avoiding the cost of implementing a false discovery.

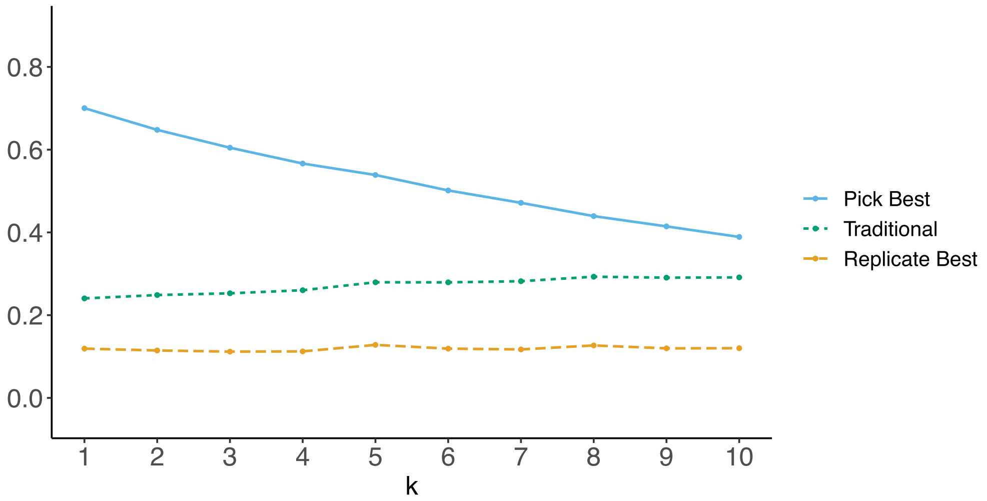
8. To What Extent Do Our Findings Generalize?

The question of how well our findings generalize has three facets, which we address in turn.

8.1. Generalizing to Other Companies

Because Optimizely is the market leader among A/B testing platforms in the United States, its customer base likely is representative of U.S. companies relying on such platforms. However, it is not representative of extremely sophisticated and intensive users of A/B testing that have developed their own internal platforms like Google, Microsoft Bing, Amazon, or

Figure 7. (Color online) False Discovery Rates for Three Testing Approaches



Booking.com. These companies often chase extremely small effects but do so using extremely large sample sizes, and it is therefore hard to speculate a priori about their typical FDR. Nevertheless, we suspect that the average FDR on Microsoft Bing (before replication) is not very different from what we estimated in our sample because their analysts report an estimate of 80% for π_0, recommend using 80% power, and note that using $\alpha = 5\%$ is common (Deng 2015, Dmitriev et al. 2017).

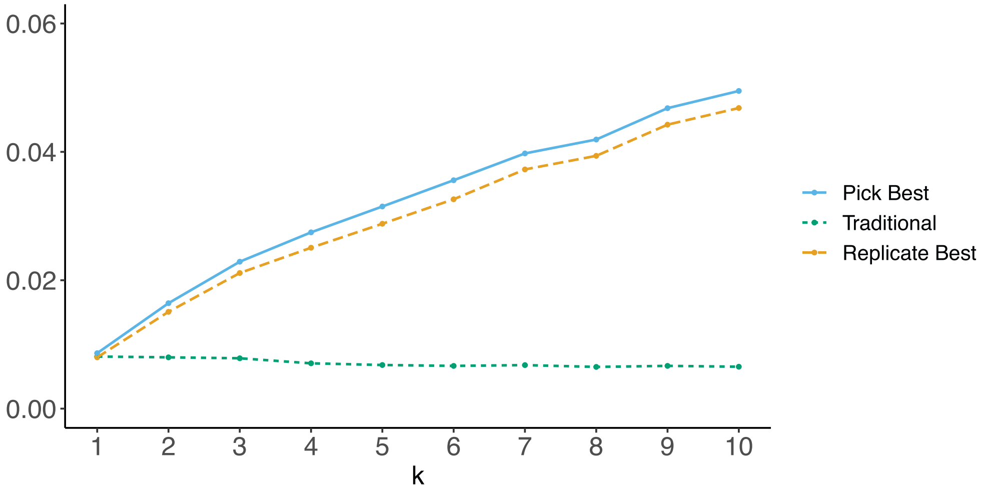
8.2. Generalizing to More Recent Experiments

As noted in Section 4.4, the experiments we analyze from 2014 are similar to those run on the same platform in 2017 and 2018 in the number of variations, run time, industry, symmetry of positive versus negative significant findings, and fraction of significant findings. Also, the pattern that the FDR decreases as an experimenter gains experience with A/B testing (Section 5.4) is likely offset by the influx of new users of A/B testing. In short, we see no reason to expect our findings to be widely off the mark of current A/B testing but very much welcome new investigations leveraging recent data.

8.3. Generalizing to More Recent Testing Practices

We calculated standard statistics for mainstream A/B tests. We doubt that two-stage designs or other methods developed recently to reduce the FDR (e.g., Deng et al. 2014) are used outside a small set of extremely sophisticated users. Also, our FDRs are expectedly

Figure 8. (Color online) Expected Gains from Experimentation (EG) for Three Testing Approaches



similar to those obtained when the statistical analysis corrects for multiple comparisons, like Optimizely's Stats Engine (Pekelis et al. 2015), other FDR-control methods, and Bayesian testing. The reason is that we declared discovery or not for each variation separately. However, to the extent that users today do not engage in "data peeking" or optional stopping or that the platform automatically corrects for it, then our FDR estimates may be higher than those currently generated. Even so, the impact of optional stopping on the FDR in the experiments we study is probably slight given the evidence in Section 6.4 and in Berman et al. (2018).

9. Conclusion

We investigate the false discovery rate in website A/B testing, i.e., the fraction of all significant results that are actually null effects. Our data consist of 4,964 effects from 2,766 experiments conducted on a commercial A/B testing platform.

Using three different methods, we find that the FDRs range between 28% and 37% for tests conducted at 10% significance and between 18% and 25% for tests at 5% significance. In other words, about one in five results significant at $p \leq .05$ is actually a true null.

These elevated FDRs stem from the high rate of true nulls (about 70%) and not from lack of statistical power on average, which we calculate to be roughly 70% at $\alpha = 0.1$ and roughly 65% at $\alpha = 0.05$. Moreover, our estimate of $\pi_0 = 0.72$ implies that even if power was increased to 100%, the FDR would not fall below 20% at $\alpha = 0.1$ and 11% at $\alpha = 0.05$. These findings support earlier suspicions that some of the frustrations with A/B testing stem from the ineffectiveness of the interventions being tested rather than from inadequacies of the method itself (Fung 2014).

More experienced experimenters tend to achieve lower FDRs, whereas the number of variations or goals in the experiment is not associated with higher or lower FDRs. The negative association between experience and FDR operates through both a lower rate of true nulls and a larger true effect size. We also find that the industries that are the heaviest users of A/B testing on this platform have lower FDRs. Our data do not allow us to determine whether the association reflects learning, larger sample sizes, or self-selection.

We use our estimates to shed some light on what changes to the basic A/B testing design might lead to a lower FDR. For the typical effect size (certainly up to the 75th percentile), tightening α and increasing sample size to keep power roughly constant hold little promise. This is likely also true for companies enjoying massive traffic but chasing very small effects. Much more promising are two-stage designs like replication. Even better is a two-stage design with multiple variations. This not only lowers the FDR but also increases the expected gain from experimentation.

Although our setting is a specific A/B testing platform, we expect our findings to be representative of many website A/B tests because our data consist of nearly 5,000 effects from 2,766 experiments, and the platform accounts for about 35% of the market for A/B testing services. Even so, several companies that use A/B tests very intensively, like Amazon, Facebook, and Microsoft, have developed their own internal platforms and are, therefore, probably not represented in our data. Such very experienced experimenters would likely enjoy lower FDRs than the typical account in our data does, were it not for the fact that they tend to chase even smaller effects than those we study. A second possible concern about the generalizability of our findings is that, after we collected our data, Optimizely and other platforms shifted from classical hypothesis testing toward FDR control methods and Bayesian hypothesis testing to account for multiple comparisons and optional stopping (e.g., Pekelis et al. 2015, Deng et al. 2016, Johari et al. 2019). However, these newer procedures by themselves do not affect the fraction of true nulls. Furthermore, our FDR calculations properly avoided multiple comparisons and are unlikely to be much affected by improper optional stopping, which is what these newer methods account for. Hence, the increasing popularity of FDR control methods and Bayesian hypothesis testing likely increases the external validity of our FDR values.

Taking a big picture perspective, one may ask what an estimate of $\pi_0 = 70\%$ implies about firm efficiency assuming that the baseline represents current practice. One possible explanation for the high fraction of true nulls is that managers struggle with generating good ideas to assess through A/B testing (Kohavi et al. 2020) either because their current practices are already close to the optimum or because the objective functions do not vary much in a wide range around the optimum (the flat maximum principle) (Sinha and Zoltners 2001). The second possible explanation is the opposite. Managers struggle to generate good ideas because of a lack of diligence, talent, or skill, suggesting a low rather than high efficiency. The third viable explanation is that experimenters engage mostly in deliberate local search through small deviations from current practice, stemming from a *kaizen* philosophy of continuous and disciplined improvement through a stream of small interventions. Given the presence of three viable explanations, one should not infer from our findings that the average firm using A/B testing is already operating near optimally.

For decision makers, our findings have four implications. (i) Have realistic expectations about your success with finding winners in your tests. Of the nearly 5,000 effects tested, the typical effect was 0, about 70% of them were probably really 0, and only 20% were significant at $\alpha = 5\%$. (ii) Have realistic expectations about your success with rolling out test winners. Even among those interventions that beat the baseline in your experiments at $\alpha = 5\%$, expect one of five to have zero effect when implemented in the field. (iii) To reduce the risk of false positives, focus on generating better ideas to test or swing for the fences by generating riskier ideas. (iv) Keep working at it. You may get better over time at reducing the fraction of false discoveries among your test results.

For analysts, our findings have two implications. (i) To lower the FDR, consider using two-stage designs with replication and A/B/n tests with multiple variations. (ii) Think about your choice of the type I error rate α. To be of greatest use to the organization, it should be informed by comparing the expected gains in effectiveness from experimentation against loss aversion, switching costs, and disappointment costs. If the goal is to maximize the expected gain in effectiveness or to minimize regret about effectiveness (Stoye 2009, Manski and Tetenov 2016, Feit and Berman 2019), then simply pick the best variation without testing, which amounts to setting $\alpha = 50\%$ in a one-sided test.

For researchers in academia, large in-house analytics teams, and platform providers, our findings raise at least three research questions. (i) What interventions are associated with a lower FDR, and how does that map into the primitives π_0, μ, and σ? (ii) Do some companies and industries have lower FDRs, and if so, why? Is it because of differences in the quality and variance of ideas or in the way they are being tested (e.g., sample size, two-stage designs)? Also, if the FDR improves with experience because of learning, what exactly is being learned, and how can it be fostered? (iii) What is the optimal FDR? Statistical decision theory does not provide an answer unless the decision maker declares their objective function. There are at least five elements that should be considered: the expected gain from experimentation EG, the setup cost of running an experiment of a given size, the switching cost of discoveries, the disappointment cost of false discoveries, and the degree of loss aversion. The optimal FDR will be a function of the weights on these components, and these weights are likely to vary across decisions and contexts.

Acknowledgments

The authors thank Optimizely for making the data available for this project and Aisling Scott and Leo Pekelis for facilitating access to the data and for helpful discussions. The authors also benefited from research assistance by Hangcheng Zhao; from comments by Elea McDonnell Feit, Raghuram Iyengar, Pete Koomen, Andrew Davies, and the review team; and from comments by audiences at the University of California, Los Angeles, the University of Chicago, Massachusetts Institute of Technology, Northeastern University, the University of Southern California, and Yale University.

Endnotes

[1] See https://vwo.com/blog/cro-industry-insights/ (accessed March 2, 2020).

[2] See https://www.datanyze.com/market-share/testing-and-optimization (accessed February 12, 2020).

[3] Even though the type II error rate is a function of α, the true effect θ, and the experimental design D, we write β instead of $\beta(\alpha, \theta, D)$ for simplicity.

[4] Although the true effect on a continuous outcome like a conversion rate cannot be exactly zero, statisticians have long considered the existence of true null effects having a strictly positive probability of occurring. This notion of true nulls is consistent with negligibly small effects, centered at zero, that require unfeasibly large samples to detect (Hodges and Lehmann 1954, Berger and Delampady 1987, Masson 2011, Deng 2015).

[5] This expression is not defined when there are no discoveries ($S = 0$). To account for such cases, one defines the FDR as $\mathbb{E}[\frac{F}{S} \mid S > 0]Pr(S > 0)$ (Benjamini and Hochberg 1995), or one uses the positive false discovery rate $(pFDR) = \mathbb{E}[\frac{F}{S} \mid S > 0]$ where the expectation conditions on having at least one positive (i.e., significant) test result (Storey 2002, 2003). When the number of effects tested m is large, $S > 0$ is almost certain for any $\alpha > 0$, and the refinement in the formal definition of the FDR and the distinction between FDR and pFDR are moot.

[6] Throughout the paper, the term experimenter refers to a unique platform account identifier (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.

[7] Thomke (2020) does not specify the level of significance and whether the tests are one-sided or two-sided tests or apply to the experiments' extreme or average variation. Consistent with Optimizely's dashboard, we assume that Thomke (2020) refers to one-sided tests at 5% significance. Finally, we assume that the numbers of Thomke (2020) pertain to the most extreme lift.

[8] Hoenig and Heisey (2001), Yuan and Maxwell (2005), Gelman and Carlin (2014), and McShane et al. (2020) express reservations about calculating power ex post using estimates from a single study, a small number of studies, or studies subject to publication bias. Because our power calculations use nearly 5,000 z scores not subject to publication bias, those reservations do not apply.

[9] The $FDR(u)$ analysis assumed that waiting one more day increases the chance of reaching significance equally among null and nonnull effects. In a sensitivity analysis, we computed an alternative $FDR(u)$, where the fraction u of new discoveries is not split between null and nonnull effects in a $\pi_0/(1-\pi_0)$ ratio but in a $\kappa/(1-\kappa)$ ratio where κ varies from 0% to 100%. Except for the case of $\alpha = 0.01$, varying κ does not affect the $FDR(u)$ values by much.

[10] Using the same sample size per arm in both steps of a preplanned replication design is not quite optimal. Our simulation therefore does not capture the full gains in FDR achievable by an optimally designed preplanned replication study.

References

Anderson ML, Magruder J (2017) Split-sample strategies for avoiding false discoveries. NBER Working Paper No. 23544, National Bureau of Economic Research, Cambridge, MA.

Azevedo EM, Deng A, Montiel Olea JL, Rao J, Weyl EG (2020) A/B testing with fat tails. *J. Political Econom.* 128(12):4614–4672.

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, Bollen KA, et al. (2018) Redefine statistical significance. *Nature Human Behav.* 2(1):6–10.

Benjamini Y, Hechtlinger Y (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.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Statist. Soc. Series B Statist. Methodology* 57(1):289–300.

Berger JO, Delampady M (1987) Testing precise hypotheses. *Statist. Sci.* 2(3):317–335.

Berman R, Feit EM (2019) Principal stratification for advertising experiments. Preprint, submitted November 19, https://arxiv.org/abs/1911.08438.

Berman R, Pekelis L, Scott A, Van den Bulte C (2018) p-Hacking and false discovery in A/B testing. Preprint, submitted July 18, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3204791.

Bhat N, Farias VF, Moallemi CC, Sinha D (2020) Near-optimal A/B testing. *Management Sci.* 66(10):4477–4495.

Blake T, Nosko C, Tadelis S (2015) Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica* 83(1):155–174.

Brodeur A, Cook N, Heyes A (2020) Methods matter: p-Hacking and publication bias in causal analysis in economics. *Amer. Econom. Rev.* 110(11):3634–3660.

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

Deng A (2015) Objective Bayesian two sample hypothesis testing for online controlled experiments. Gangemi A, Leonardi S, Panconesi A, eds. *Proc. 24th Internat. Conf. World Wide Web* (Association for Computing Machinery, New York), 923–928.

Deng A, Li T, Guo Y (2014) Statistical inference in two-stage online controlled experiments with treatment selection and validation. Chung C-W, ed. *Proc. 23rd Internat. Conf. World Wide Web* (Association for Computing Machinery, New York) 609–618.

Deng A, Lu J, Chen S (2016) Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing. Osmar RZ, Matwin S, eds. *2016 IEEE Internat. Conf. Data Sci. Advanced Analytics (DSAA)* (IEEE, Piscataway, NJ), 243–252.

Dmitriev P, Gupta S, Kim DW, Vaz G (2017) A dirty dozen: Twelve common metric interpretation pitfalls in online controlled experiments. Matwin S, Yu S, Farooq F, eds. *Proc. 23rd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 1427–1436.

Efron B (2012) *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (Cambridge University Press, Cambridge, United Kingdom).

Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* 96 (456):1151–1160.

Feit EM, Berman R (2019) Test & roll: Profit-maximizing A/B tests. *Marketing Sci.* 38(6):1038–1058.

Fung K (2014) Yes, A/B testing is still necessary. *Harvard Bus. Rev.* (December 10), https://hbr.org/2014/12/yes-ab-testing-is-still-necessary.

Gelman A, Carlin J (2014) Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspect. Psych. Sci.* 9(6):641–651.

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

Gordon BR, Zettelmeyer F, Bhargava N, Chapsky D (2019) A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Sci.* 38(2):193–225.

Gronau QF, Duizer M, Bakker M, Wagenmakers E-J (2017) Bayesian mixture modeling of significant p values: A meta-analytic method to estimate the degree of contamination from H_0. *J. Experiment. Psych. General* 146(9):1223–1233.

Hodges J, Lehmann E (1954) Testing the approximate validity of statistical hypotheses. *J. Royal Statist. Soc. Series B Statist. Methodology* 16(2):261–268.

Hoenig JM, Heisey DM (2001) The abuse of power: The pervasive fallacy of power calculations for data analysis. *Amer. Statist.* 55(1):19–24.

Hung HJ, O'Neill RT, Bauer P, Kohne K (1997) The behavior of the p-value when the alternative hypothesis is true. *Biometrics* 53 (1):11–22.

Ioannidis JP (2005) Why most published research findings are false. *PLoS Medicine* 2(8):e124.

Johari R, Pekelis L, Walsh DJ (2019) Always valid inference: Bringing sequential analysis to A/B testing. Preprint, submitted July 16, https://arxiv.org/abs/1512.04922.

Johari R, Koomen P, Pekelis L, Walsh D (2017) Peeking at A/B tests: Why it matters, and what to do about it. Matwin S, Yu S, Farooq F, eds. *Proc. 23rd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 1517–1525.

Johnson G, Lewis RA, Nubbemeyer EI (2017a) The online display ad effectiveness funnel & carryover: Lessons from 432 field experiments. Preprint, submitted October 2, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2701578.

Johnson V, Payne RD, Wang T, Asher A, Mandal S (2017b) On the reproducibility of psychological science. *J. Amer. Statist. Assoc.* 112(517):1–10.

Kohavi R, Tang D, Xu Y (2020) *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing* (Cambridge University Press, Cambridge, United Kingdom).

Korthauer K, Kimes PK, Duvallet C, Reyes A, Subramanian A, Teng M, Shukla C, Alm EJ, Hicks SC (2019) A practical guide to methods controlling false discoveries in computational biology. *Genome Biology* 20(1):118.

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

Lewis RA, Rao JM (2015) The unfavorable economics of measuring the returns to advertising. *Quart. J. Econom.* 130(4): 1941–1973.

Lu M, Stephens M (2019) Empirical Bayes estimation of normal means, accounting for uncertainty in estimated standard errors. Preprint, submitted January 30, https://arxiv.org/abs/1901.10679.

Maniadis Z, Tufano F, List JA (2014) One swallow doesn't make a summer: New evidence on anchoring effects. *Amer. Econom. Rev.* 104(1):277–290.

Manski CF, Tetenov A (2016) Sufficient trial size to inform clinical practice. *Proc. Natl. Acad. Sci. USA* 113(38):10518–10523.

Masson ME (2011) A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behav. Res. Methods* 43(3): 679–690.

McShane BB, Böckenholt U, Hansen KT (2020) Average power: A cautionary note. *Adv. Methods Practices Psych. Sci.* 3(2):185–199.

Pekelis L, Walsh D, Johari R (2015) The new stats engine. Technical report, Optimizely, San Francisco.

Pocock SJ (1977) Group sequential methods in the design and analysis of clinical trials. *Biometrika* 64(2):191–199.

Pounds S, Cheng C (2006) Robust estimation of the false discovery rate. *Bioinformatics* 22(16):1979–1987.

Sarkar SK, Chen J, Guo W (2013) Multiple testing in a two-stage adaptive design with combination tests controlling FDR. *J. Amer. Statist. Assoc.* 108(504):1385–1401.

Scott JG, Kelly RC, Smith MA, Zhou P, Kass RE (2015) False discovery rate regression: An application to neural synchrony detection in primary visual cortex. *J. Amer. Statist. Assoc.* 110(510):459–471.

Sinha P, Zoltners AA (2001) Sales-force decision models: Insights from 25 years of implementation. *Interfaces* 31(3 supplement):S8–S44.

Storey JD (2002) A direct approach to false discovery rates. *J. Royal Statist. Soc. Series B Statist. Methodology* 64(3):479–498.

Storey JD (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* 31(6):2013–2035.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* 100(16):9440–9445.

Storey JD, Bass AJ, Dabney A, Robinson D (2019) qvalue: Q-value estimation for false discovery rate control. R package version 2.14.1.

Stoye J (2009) Minimax regret treatment choice with finite samples. *J. Econometrics* 151(1):70–81.

Thomke SH (2020) *Experimentation Works: The Surprising Power of Business Experiments* (Harvard Business Press, Boston).

Wald A (1945) Sequential tests of statistical hypotheses. *Ann. Math. Statist.* 16(2):117–186.

Yuan K-H, Maxwell S (2005) On the post hoc power in testing mean differences. *J. Ed. Behav. Statist.* 30(2):141–167.

Zehetmayer S, Posch M (2012) False discovery rate control in two-stage designs. *BMC Bioinformatics* 13(1):81.

Zehetmayer S, Bauer P, Posch M (2005) Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics* 21(19):3771–3777.