# Statistical Challenges in Online Controlled Experiments: A Review of A/B Testing Methodology

Nicholas Larsen[*], Jonathan Stallrich[*], Srijan Sengupta[*], Alex Deng[†], Ron Kohavi, Nathaniel T. Stevens[‡]

**Abstract**

The rise of internet-based services and products in the late 1990's brought about an unprecedented opportunity for online businesses to engage in large scale data-driven decision making. Over the past two decades, organizations such as Airbnb, Alibaba, Amazon, Baidu, Booking.com, Alphabet's Google, LinkedIn, Lyft, Meta's Facebook, Microsoft, Netflix, Twitter, Uber, and Yandex have invested tremendous resources in *online controlled experiments* (OCEs) to assess the impact of innovation on their customers and businesses. Running OCEs at scale has presented a host of challenges requiring solutions from many domains. In this paper we review challenges that require new statistical methodologies to address them. In particular, we discuss the practice and culture of online experimentation, as well as its statistics literature, placing the current methodologies within their relevant statistical lineages and providing illustrative examples of OCE applications. Our goal is to raise academic statisticians' awareness of these new research opportunities to increase collaboration between academia and the online industry.

**Keywords:** Online controlled experiments, A/B testing, literature review, randomized controlled trials, treatment effect estimation

[*]North Carolina State University, Department of Statistics
[†]Airbnb
[‡]University of Waterloo, Department of Statistics & Actuarial Science, nstevens@uwaterloo.ca

# 1 Introduction

## 1.1 Background

It's estimated that in 2021 globally, 4.95 billion people (62.5% of the world's population) used the internet, each engaging with it on average 7 hours per day, and in aggregate spending over $5 trillion USD on consumer goods, travel and tourism, digital media, online food delivery, and health-related products and services (Kemp, 2022). In 2022, e-commerce is predicted to account for 21% of all commerce, and by 2025 that number is expected to grow to nearly 25% (Keenan, 2022). Given this scale of internet use, it is unsurprising that the optimization of online products and services is of great interest to online businesses and online components of traditional brick-and-mortar businesses.

*Online controlled experiments* (OCEs), digital versions of randomized controlled trials, or RCTs (Box et al., 2005), provide a means to do this; OCEs seek to use user-generated data to test and improve internet-based products and services. Informally referred to as *A/B tests*, OCEs are an indispensable tool for major technology companies when it comes to maximizing revenue and optimizing the user experience (Luca and Bazerman, 2021). Industry giants run hundreds of experiments on millions of users every day, testing changes along multiple axes including: websites, services, and installed software; desktop and mobile devices; front- and back-end product features; personalization and recommendations; and monetization strategies. With OCEs, the causal impact of such changes–whether it be positive, negative, or zero–can be measured. While most positive changes are small, and improvement is incremental (Bojinov and Gupta, 2022), results from OCEs have the potential to be incredibly lucrative. Google's famous "41 shades of blue" experiment is a classic example of an OCE that translated into a $200 million USD increase in annual revenue (Hern, 2014); Amazon used insights from an OCE to move credit card offers from the homepage to the checkout page, resulting in tens of millions of USD in profit annually (Kohavi and Thomke, 2017); Bing deployed an A/B test for ad displays that resulted in $100 million USD

of additional revenue in the U.S. alone (Kohavi et al., 2020). Even though such million-dollar ideas are relatively rare, the net gains from OCEs have been so profound that many organizations have completely overhauled their business models, with experimentation at the epicenter (Thomke, 2020). For instance, Netflix attributes its membership growth from 2 countries to over 190 in the span of just 6 years to its adoption of online controlled experimentation (Urban et al., 2016), and Duolingo's 2022 Q2 shareholder letter attributes their growth to an "A/B test everything" mentality (Von Ahn, 2022). The document even includes a description of their A/B testing process and several examples of how the product as evolved through experimentation.

Organizations that have accepted OCEs as standard practice generally adopt a so-called "culture of experimentation," which is rooted in three tenets (Kohavi et al., 2013): (1) the organization wants to make data-driven decisions, (2) the organization is willing to invest in the people and infrastructure needed to run trustworthy experiments, and (3) the organization recognizes that it is poor at assessing the value of ideas. Generally, more than 50% of ideas fail to generate meaningful improvements (Kohavi et al., 2020). Carefully executed experiments therefore provide a trustworthy, data-driven means to determine which ideas improve key metrics, which hurt, and which have no detectable impact, allowing the organization to invest in those that work, while pivoting to avoid the others. Within this culture, the attitude of "*more, better, faster*" is prevalent (Tang et al., 2010); organizations strive to increase the number of experiments so that all changes are properly evaluated; invalid experiments and harmful combinations of variants are straightforward to identify; and deployment, run time, and analysis occur within a relatively short period of time.

Compared to physical RCTs (in e.g., agriculture, manufacturing, pharmaceutical development), the cost incurred to design and run an OCE is low, even negligible for organizations with expertise in software development and statistics. Consequently, practitioners are able to run large numbers of experiments with potentially enormous sample sizes. In the case of large tech organizations, the combination of new features and modifications can result in billions

of different versions of a given product (Kohavi et al., 2013), with hundreds of thousands of users randomized to hundreds of experiments at a time (Gupta et al., 2019). Companies performing OCEs at this scale typically develop in-house software (i.e., an *experimentation platform*) to automate the experimentation process, such as randomizing users, collecting data, managing concurrent experiments, and generating analysis reports (Kohlmeier, 2022; Tang et al., 2010; Ivaniuk, 2020; Fabijan et al., 2018). See Visser (2020) for a catalogue of in-house experimentation platforms developed by several prominent tech companies. Smaller companies, on the other hand, tend not to handle these tasks themselves, and instead opt for third-party vendors that specialize in setting up, deploying, and analyzing OCEs. Popular vendors as of 2022 include Optimizely, Google Optimize, AB Tasty, VWO, Split, and Statsig (Kohavi, 2022). In all cases, this level of automation necessitates data quality checks like A/A tests and sample ratio mismatch (SRM) tests to establish trust in the experimentation platform. (For further discussion of these practices and challenges, see Chapters 19 and 21 of Kohavi et al., 2020, and the introduction in Lindon and Malek, 2020.)

In this online setting, with the culture of testing as many ideas as possible, as quickly as possible, novel practical issues and modern challenges abound (see, e.g., Gupta et al. (2019) and Bojinov and Gupta (2022) for nontechnical discussions, and Georgiev (2019) for a technical primer). The context in which OCEs operate departs markedly from the original applications for which traditional RCTs were developed nearly a century ago; understanding this context is vital for developing relevant methodology for OCEs. For statisticians, online controlled experimentation provides a host of new opportunities for methodological and theoretical development. New approaches that fit the nuances of OCE applications are in high demand, with the majority of cutting-edge research spearheaded by those in industry. The purpose of this paper, therefore, is to review the statistical methodology associated with OCEs, summarize its accompanying literature, and provide an overview of open statistical problems with the intent to increase academic statisticians' awareness of these research opportunities and to bridge the gap between academia and the online industry.

## 1.2 The General Framework

Here we introduce the notation and key terms that will be used throughout this review and we describe the basic statistical framework for OCEs. It is useful to note that as a field, online experimentation has developed disparately across industries and domains, thus there are no unifying standards with respect to methodological approach and notation; even the term "controlled experiment" goes by different names depending on the organization: "flights" at Microsoft, "bucket tests" at Yahoo, "field experiments" at Facebook, and "1% experiments/click evaluations" at Google. Standard conventions would bring useful unification to this field. The following notation largely draws from traditional randomized controlled trials and causal inference literature, and is intended to help unify much of the OCE literature.

Let $T$ be the number of variants (also known as buckets, arms, and splits) that compose the experiment. Ordinarily one of these variants is a control against which all other variants are compared. Unless explicitly stated, we shall assume for the rest of this review a standard treatment-versus-control setup, in which case $T = 2$. While multi-variant ($T > 2$) experiments exist in this space (they're colloquially referred to as "A/B/n tests"), we focus on the $T = 2$ "A/B test" for pedagogical reasons; even with $T > 2$ variants, determining which is optimal typically reduces to a pairwise comparison between each treatment and the control.

In such experiments, $n$ experimental units (e.g., users, cookies, sessions, etc.) are often randomized in real time to one of the variants, and a response observation $Y_i$ is collected for each $i = 1, \ldots, n$. It's important to note that these response observations are typically themselves aggregates of more granular raw event data (Boucher et al., 2020). For instance, consider the response variable *number of clicks per user* which may be a count per user aggregated across sessions and/or pages. Interest then lies in optimizing some *metric*, which is a numerical summary of the response. Extending the previous example, interest may lie in maximizing the *average* number of clicks per user. Such metrics are often, but not always, averages. In some contexts, quantile or double-average metrics may be more suitable. We discuss such applications in more detail in Section 7.

For simplicity of exposition, we've described a situation with one metric and hence one response variable. However, in practice there may be hundreds (even thousands) of metrics computed, many of which are used for debugging, some of which may be organizational *guardrail* metrics that the experimenters wish to avoid negatively impacting, and a small number of which compose the *overall evaluation criterion (OEC)* which is to be optimized. In general, defining and selecting metrics (as well as their corresponding randomization and analysis units) are key components of OCEs and we direct the reader to Crook et al. (2009), Deng and Shi (2016), Dmitriev et al. (2017), and Kohavi et al. (2020) for further discussion.

When the metric is an average, the primary goal of the experiment is to estimate the *Average Treatment Effect (ATE)*; the difference between the average outcome when the treatment is applied globally and when the control is applied globally. Within the potential outcomes framework (Rosenbaum and Rubin, 1983), $Y_i(0)$ represents unit $i$'s response in the hypothetical scenario where $i$ receives the control, and $Y_i(1)$ is the potential response when unit $i$ receives the treatment. Letting $W_i$ denote the binary treatment indicator for unit $i$, and given a particular treatment assignment to all experimental units $\mathbf{W} = \big(W_1, \ldots, W_n\big)'$, the expected outcome is $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i(W_i)\right] = \mu(\mathbf{W})$, and the ATE is therefore given by

$$\begin{aligned} \tau &= \mu(\mathbf{1}) - \mu(\mathbf{0}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i(1) - Y_i(0)]. \end{aligned} \tag{1}$$

In reality, $i$ can only be assigned to a single variant at a time, thus one cannot directly observe both $(Y_i(0), Y_i(1))$ and so the ATE is typically estimated with the difference-of-group-means estimator,

$$\hat{\tau} = \frac{1}{n_1}\sum_{\{i:W_i=1\}} Y_i - \frac{1}{n_0}\sum_{\{i:W_i=0\}} Y_i, \tag{2}$$

where $n_0$ and $n_1$ are respectively the sizes of the control and treatment groups such that $n_0 + n_1 = n$. In practice it is also common to define the treatment effect as a relative percent, often referred to as *lift*, since it is easier to interpret and it is more stable (over experiment

duration, for example).

Statistical significance is the most common mechanism by which a given treatment's effectiveness is affirmed in an A/B test. Analyses of A/B tests are therefore most often carried out via two-sample hypothesis tests for $\tau$ with standard test statistics of the form $\hat{\tau}/\hat{\sigma}_{\hat{\tau}}$. Such analyses, and the designs that generate data for them, commonly assume that the response of each user does not depend on other users' treatment assignments (the *Stable Unit Treatment Value Assumption*, or SUTVA). SUTVA is a reasonable assumption for many scenarios; however in Section 6 we discuss OCE settings where the assumption is violated and alternative methodologies are necessary. In many scenarios, sample sizes are large enough to confidently exploit the central limit theorem, permitting the use of the standard normal null distribution. There are, however, scenarios in which only a fraction of the user base is experimented on and asymptotic normality cannot be assumed. Such scenarios are discussed in Section 2.2. Given the heavy reliance on p-values it's important to acknowledge that the reproducibility crisis stemming from the misuse of hypothesis tests also plagues OCEs; p-value misinterpretation and problematic practices regularly lead to increased false-positive rates (Berman and Van den Bulte, 2021; Kohavi et al., 2022). This is an area of ongoing practical and methodological concern in many fields, including online experiments.

## 1.3 Roadmap

With this context and foundation laid, we now review the statistical research in this area and discuss the many open problems. The article proceeds as follows. Section 2 discusses techniques for improving experimental power – a critical issue despite the relatively large sample sizes in OCEs. Sections 3 and 4 respectively present literature regarding the challenges of estimating heterogeneous and long-term treatment effects. Section 5 discusses the problem of optional stopping and how sequential testing methods have been adapted to run online experiments. All of these sections presume SUTVA holds; we summarize the literature that explores violations of this assumption in Section 6. A brief discussion of additional

topics is presented in Section 7 and we conclude the review with a call to action for further collaboration between academia and OCE practitioners in Section 8.

## 2 Sensitivity and Small Treatment Effects

**Motivating Example:** *Suppose an e-commerce website observes that 5% of their visitors make a purchase and the average purchase is \$25 with a standard deviation of \$6 during a one-week experiment period. Therefore the "average" visitor spends \$1.25. Suppose also the company's annual revenue is \$20 million, and gains or losses of \$1 million (5%) are material. If the company wishes to run an experiment and detect a 5% change in revenue with 80% power at a 5% significance level, a rough sample size calculation indicates they need* $n_0 = n_1 = 16\sigma^2/\delta^2 = (16 \times 6^2)/(1.25 \times 0.05)^2 = 147,456$ *users per variant. This is reasonable for a small startup. However, suppose now that the company's annual revenue is \$50 billion, with gains or losses of \$10 million (0.02%) of interest. An experiment designed to detect a 0.02% change in revenue requires* $n_0 = n_1 = 16\sigma^2/\delta^2 = (16 \times 6^2)/(1.25 \times 0.0002)^2 = 9.2$ *billion users per variant, i.e., 18.4 billion users in a single week. The human population on Earth is about 8 billion at the time of writing, so it's impossible for this company to detect changes that would lose them \$10 million/year.*

Many leading organizations at the forefront of online controlled experimentation have user populations numbering in the hundreds of millions, if not billions. However, the sentiment that OCEs do not suffer from inadequate sample sizes is misconceived (Tang et al., 2010). Given the fundamental relationship between sample size and an experiment's ability to detect true, nonzero treatment effects (namely its power), a key challenge facing even the largest of organizations is designing adequately powered experiments. A naive solution would be to simply extend the experiment's duration, thereby increasing the number of users. However, as we elaborate upon in Section 4, this practice is ill-advised. Instead, it is better practice to employ a tactic that is tailored to the cause of insufficient power, which is generally one

of three things.

First, the treatment impacts the entire user population and the effect is roughly homogeneous, but very small in magnitude. As illustrated in the opening example, even a fraction of a percent-change can translate to millions of dollars in revenue. We discuss the literature around this issue in Section 2.1. Second, many experiments test features that do not affect all users, making the treatment effect highly attenuated (Section 2.2). Third, the treatment effects on known subpopulations are of interest, where sample sizes are smaller by definition (we defer this discussion to Section 3). In general, research regarding improving experimental power for OCEs tends to focus on boosting *sensitivity*, either by directly reducing the variance of $Y_i$ or by improving the bias and variance of estimators for $\tau$. The aforementioned subsections provide an overview of common methodology in this area. Although not covered in detail here, there is also growing interest in altering the *design* of the experiment to boost sensitivity. In some contexts, interleaving, for instance, is an effective means to decrease necessary sample sizes, particularly when query-related metrics are of interest (Radlinski and Craswell, 2013; Parks et al., 2017; Zhang et al., 2022). While specific methods of combating inadequate power are reviewed in this section, we encourage the reader to keep in mind that the issue of adequate power applies to all the challenges subsequently discussed in this review.

## 2.1 Transforming $Y$, Method of Control Variates, and Stratified Sampling

In order to improve sensitivity, a common approach is to transform $Y$ into $Y^*$ of lower variance which, all else being equal, translates to a lower variance estimator of $\tau$. In online experiments there can be dozens, even hundreds of metrics of potential interest, many with different properties that make it all but impossible to identify a "one size fits all" transformation. Much work has been devoted to documenting metric behavior and discussing techniques for metric definition. Kohavi et al. (2014) describe several examples of non-intuitive met-

ric behavior and other peculiarities, illustrating the benefits of identifying skewed metrics and capping (truncating) them to improve sensitivity. Other transformations for improving the sensitivity of $Y$ include binarizing count metrics and revenue. Deng and Shi (2016) define *directionality* (consistent behavior in one direction for positive treatment effects, in the opposite direction for negative effects) as an important feature when choosing metrics, suggesting that one should leverage prior experiments to compile a corpus of good metrics and to evaluate sensitivity and directionality with Bayesian priors. Deng and Shi (2016) also propose aggregating metrics in the form of a weighted linear combination, which is adopted and expanded upon in Kharitonov et al. (2017). They frame finding sensitive combinations of metrics as a machine learning problem, incorporating both labeled and unlabeled data from past experiments. In Drutsa et al. (2015a), features are extracted from data while the experiment is running and used to forecast metrics over a hypothetical post-experiment period. The authors also note their methodology may be applied to long-term effect estimation using statistical surrogacy, which we further discuss in Section 4.

In addition to transformations of $Y$, a popular approach is to define an efficient, mean-zero augmented estimator of $\tau$ using the method of control variates (Courthoud, 2022; Sexauer, 2022; Sharma, 2022). Briefly, this method assumes, in addition to i.i.d $\{Y_i\}_{i=1}^n$, the availability of independent observations of a covariate, $\{X_i\}_{i=1}^n$, such that $\mathbb{E}[X_i] = \mu_x$. Often, these covariate measurements are collected from prior logs or experiments. Let $Y_i^* = Y_i - \theta(X_i - \mu_x)$, then $Var(Y_i^*) = Var(Y_i) + \theta^2 Var(X_i) - 2\theta Cov(Y_i, X_i)$ is minimized with respect to $\theta$ using the OLS solution $\frac{Cov(Y_i, X_i)}{Var(X_i)}$. Putting this together in the context of sample means gives $Var(\overline{Y}^*) = (1 - \rho^2) Var(\overline{Y}) \leq Var(\overline{Y})$, where $\rho = Corr(Y_i, X_i)$. Thus, an ATE estimator that uses the difference of treatment and control means of $Y_i^*$ tends to have lower variance than the traditional $\hat{\tau}$, particularly when $X_i$ is strongly correlated with $Y_i$. For OCEs, this technique is referred to as CUPED (Controlled experiments Utilizing Pre-Experiment Data) and was first proposed by Deng et al. (2013). The authors empirically demonstrate that an effective covariate choice is the same variable $Y_i$ but collected during a pre-experiment

period ($X_i \equiv Y_i^{\text{pre}}$). Such a choice can drastically increase sensitivity and thereby reduce time to statistical significance in determining $H_1 : \tau \neq 0$. The authors also demonstrate that $\mu_x$ need not be known when $X_i$ is uncorrelated with $W_i$ and they also emphasize that despite resembling ANCOVA, CUPED does not require any linear model assumptions and can be treated as efficiency augmentation as in semi-parametric estimation (Tsiatis, 2006). Consequently, CUPED has become a standard tool for many practitioners, although it is important to note that it can only be applied to users for which prior information exists (Gupta et al., 2019; Hopkins, 2020; Drutsa et al., 2015b; Jackson, 2018; Sharma, 2021).

A key open question with respect to CUPED applications concerns the situation when the covariate alone is not sufficiently correlated with the response. An approach that shows promise employs synthetic controls, where one identifies a similar population without treatment exposure to use as covariates for modeling $Y$ (Zhang et al., 2021). Another technique is to take advantage of a phenomenon that occurs in online experiments known as "triggering" (Deng et al., 2021b), which we further discuss in Section 2.2. Further research with respect to the interplay between CUPED and other standard variance reduction techniques is also of interest. Xie and Aurisset (2016) apply CUPED to large-scale A/B tests for a subscription streaming service, and Liou and Taylor (2020) compare CUPED against variance-weighted estimators on a social media platform, finding that an aggregation of the two methods outperformed either individually. Deng et al. (2013) note that CUPED also permits nonlinear adjustments to the response variable. Following this, Poyarkov et al. (2016) develop an approach that assumes each user has a response $Y$ and a set of features $\mathbf{F} \in \mathbb{R}^p$ independent of treatment assignment. Let $Y = f(\mathbf{F})$, where $f$ is an unknown, non-parametric function that is estimated with machine learning. Following the general idea of control variates, the covariate is chosen to be the predicted outcomes of $\hat{f}$. Poyarkov et al. (2016) then use $Y^* = Y - \hat{f}(\mathbf{F})$ as the primary metric for estimating the ATE, noting an increase in sensitivity compared to traditional A/B tests.

Closely related to the method of control variates/CUPED is stratified sampling. Assume

there exist $K$ strata dividing the population $\Omega$, where every stratum has mean and variance $(\mu_k,\ \sigma_k^2)$, and each unit $i$ falls into the $k^{\text{th}}$ strata with unknown probability $w_k$ such that $\sum_{k=1}^{K} w_k = 1$. With data obtained via stratified sampling, it is well-known that one may construct an unbiased, weighted estimator of $\tau$ that has smaller variance than the standard difference-of-means estimator, presuming one has correctly estimated $w_k$ and identified stratum that are correlated with $Y$ (Acharya et al., 2013). As noted in Deng et al. (2013) and Xie and Aurisset (2016), many organizations have access to large amounts of data, which can simplify the process of identifying meaningful strata. However, estimating $w_k$ is not straightforward, and the real-time nature of online experiments as well as the physical infrastructure of experimentation platforms also hinder accurate implementation of stratified random sampling. The primary challenge is to maintain equal representation of the strata while users are randomized to treatment and control. Xie and Aurisset (2016) propose a novel stratified sampling technique that involves defining one queue $q$ for each strata $k$. Each $q$ consists of multiple segments of fixed length. Depending on their strata, users are first assigned to a slot within a segment, then treatments are randomized within each segment. Consequently, balanced allocation is only guaranteed within a segment. Moreover, if multiple machines each have their own $q$ for strata $k$, as is the case in many large experimentation platforms, balanced randomization is even more difficult to achieve. Deng et al. (2013) show that CUPED is equivalent to stratified random sampling when the control variate is categorical, and is considered a post-experiment workaround for the practical difficulties of implementing stratified sampling in real-time. Xie and Aurisset (2016) compare their stratified sampling technique to CUPED, finding that CUPED prevails in terms of variance reduction. Practitioners continue to be interested in methods for stratified sampling with the aim of variance reduction, as well as identifying such strata in order to detect bugs or potential areas for targeted optimization.

## 2.2 Triggered Analysis

**Motivating Example:** *Suppose engineers are testing a change made on an e-commerce website's checkout page. Users in the experiment who never interact with this checkout page are not impacted by the experiment and so their treatment effect is zero. Many such users will increase noise and dilute the treatment effect. So, sensitivity may be increased by analyzing only the users who could have been impacted by the experiment; those that were triggered into the analysis. Although this reduces sample size, the treatment effect among the triggered users is undiluted and therefore higher and easier to detect.*

*Triggered analysis* broadly refers to an OCE analysis where one only considers users who have the potential of being impacted by an experiment, excluding those who would not be effected by the proposed variant (Deng et al., 2021b; Kohavi et al., 2009; Kohavi, 2012; Xu et al., 2018). Users are said to have *triggered* the experiment when, after being randomized at an earlier stage, they exhibit behavior that results in direct exposure to their assigned variant. In the checkout example above, users may be assigned to treatment or control upon entering the homepage, but in order to actually experience the designated variant, the users must first navigate to the checkout page. Key analysis challenges include: (1) generalizing the results from the triggered users to a broader population, and (2) reducing the variance of $\tau$ estimators to offset the smaller sample sizes that result from triggering. For an in-depth discussion of triggering case-studies, including the example above, see Chapter 20 of Kohavi et al. (2020).

### 2.2.1 Review

Let $\Omega$ be the overall user population and $\Theta \subset \Omega$ the population of users who could be effected by the treatment. A given user is determined to belong to $\Theta$ via techniques such as conditional checks or counterfactual logging (Kohavi et al., 2020; Deng et al., 2021b). If $\Theta$ comprises only a modest fraction of $\Omega$, (i.e., $\frac{|\Theta|}{|\Omega|} \leq 0.2$, for instance), an experiment that samples data from the entire population could be severely under-powered, particularly when

effect sizes are small (Kohavi et al., 2009). To mitigate this issue, practitioners focus analysis only on triggered users. The difference-of-means estimator $\hat{\tau}_\Theta$ is an unbiased estimator for the ATE of the triggered population, $\tau_\Theta$, under standard assumptions. However, $\tau_\Theta$ is typically larger than the population-level $\tau_\Omega$ and the corresponding estimator generally has greater variance. In practice, the goal is to obtain statistically significant evidence as to whether or not $\tau_\Omega = 0$. The process of estimating $\tau_\Omega$ with $\hat{\tau}_\Theta$ is referred to as *estimating the diluted treatment effect*. This section focuses on commonly used methods for estimating the diluted treatment effect, including discussion regarding their current limitations and how they have been addressed in the literature.

Most triggered analyses fall under the following framework. Assume a random sample of $N$ units, $n$ of which are triggered. Each user $i$ interacts with the website on multiple separate events. During each event, $i$ may or may not trigger (e.g., $i$ may interact with the checkout page during one event, but not the other). The most common analysis technique is the *user-trigger analysis*, which incorporates all events beginning with the first event where $i$ triggered. Such analyses are quite popular as they do not require any assumptions regarding the treatment effect, and are amenable to common user-level metrics. Chen et al. (2018) utilize the user-trigger framework to illustrate the benefits of triggered analyses in terms of power gains and variance reduction, as well as to highlight the types of biases that may occur under such approaches. The *session-trigger analysis* is another approach that groups events into "sessions" and only keeps sessions that contain trigger events for analysis. While Deng and Hu (2015) note that estimates from session-triggered analyses do tend to have lower variance than user-trigger analyses, the treatment effect must be zero in the nontriggered sessions in order for this approach to be valid. While perhaps true in some cases, generally this assumption is difficult to verify for most applications.

One approach for estimating the diluted treatment effect is to derive $\tau_\Omega$ in terms of $\tau_\Theta$, producing so-called "diluted formulas". For additive metrics, $Y_i = Y_{i,t} + Y_{i,u}$, where $Y_{i,t}$ is an outcome when $i$ is triggered and $Y_{i,u}$ for when $i$ is untriggered, it can be shown

that the diluted treatment effect is the average treatment effect on triggered users weighted by the proportion of triggered users, i.e., $\tau_\Omega = \tau_\Theta \times \frac{n}{N}$. Note that this expression only applies when, for $i \in \Theta$, there is no treatment effect on the sessions where $i$ is untriggered, i.e., $Y_{i,u}(1) - Y_{i,u}(0) = 0$. In other words, this expression is only for valid session-trigger analyses. With ratio metrics, $Y_i = \frac{a_i}{b_i}$, if there is no treatment effect for the denominator term, $b_i = b_i(1) = b_i(0)$, and the rate at which users are triggered into the experiment is independent of $\tau_\Theta$, then the diluted formula is $\tau_\Omega = \tau_\Theta \times \frac{n}{N} \times \bar{r}$, where $\bar{r}$ is the average trigger rate as a function of $b_i$. Further details for these derivations may be found in Deng and Hu (2015). While these formulas are certainly helpful in illustrating the connection between $\tau_\Theta$ and $\tau_\Omega$, they are restrictive because their underlying assumptions are not necessarily realistic and closed-form expressions only exist for special cases. As noted by Deng and Hu (2015), the trigger rates are rarely independent of the triggered treatment effect. Users who visit a website frequently will have higher trigger rates and tend to have a larger treatment effect than less-frequent users (see Wang et al. (2019)).

The above formulas are but one solution to the question of estimating $\tau_\Omega$ – they still do not address the issue of low power that typically afflicts triggered analyses. Deng and Hu (2015) and Deng et al. (2021b) simultaneously address both issues by formalizing the connection between all diluted formulas and variance reduction. Under the assumption that there is no treatment effect when users are untriggered, Deng and Hu (2015) apply the CUPED framework (Section 2.1) by augmenting $\hat{\tau}_\Theta$ with mean-zero data from the trigger complement group. The authors show that the resulting augmented estimator is unbiased for $\tau_\Omega$, can achieve appreciable variance reduction, and applies to metrics of any form. Deng et al. (2021b) extend this application of CUPED to one-sided triggering, a type of one-sided noncompliance where only the triggering status of the treatment group is observed.

Compared to other challenges in OCEs, the literature for triggering is, at present, rather sparse. Consequently, there are still many areas open for further research. The discussed methodologies for estimating the diluted treatment effect each depend on assumptions that

may be too restrictive in certain applications. An additional challenge concerns bias of standard ATE estimators induced by triggering. Chen et al. (2018) identify a special type of bias that occurs when a user triggers on day $k$, but not day $k+1$. Other types of bias, as well as the questions of how to define the randomization unit (user, session, or webpage) and how and when to aggregate data into sessions, remain open for exploration. Recent work in Deng et al. (2021b) also suggests that exploring noncompliance and other similar concepts from the causal inference literature (such as principle stratification) with respect to triggering may be an area for potentially interesting future development.

# 3 Heterogeneous Treatment Effects

**Motivating Example:** *Suppose an online ad provider wishes to determine the impact of changing from static textual ads to short video ads on website traffic. For the treatment group, website traffic appears to have increased uniformly except among Safari users. Consequently, the ad team wishes to estimate the treatment effect at the browser level. Likewise, after observing an improvement in user engagement metrics, the ad team may want to perform a post-hoc analysis to determine if this increase is roughly the same for all users or perhaps concentrated within certain user segments (such as those defined by market/country, user activity level, device/platform type, and time).*

Treatment effects on subgroups that differ from the population-level ATE are known as *heterogeneous treatment effects (HTE)* and are commonly of interest in OCEs. Identifying and interpreting such heterogeneity is vitally important for business applications. For example, practitioners may be interested in estimating the treatment effect for different devices or browser types, or for users of different ages, or users living in different parts of the world. Identifying and estimating HTEs is also of concern for those wishing to develop personalized experiences, or detect bugs, or interactions with other experiments. Three key challenges are associated with estimating HTEs: (1) small treatment effects (see Section 2) often make

online studies under-powered, resulting in high false positive rates for subgroup effects; (2) testing for HTEs tends to risk inflated Type I error rates due to multiple comparisons; (3) users are generally not randomized to the subgroups under comparison, so the usual tension between correlation and causation manifests itself here. Below we review existing methodologies that are commonly used in the context of OCEs to address this problem.

## 3.1 Review

Heterogeneous treatment effects have a rich history in statistical theory and application (Robinson, 1988; Athey and Imbens, 2016; Wager and Athey, 2018; Zhao et al., 2012; Tran and Zheleva, 2019; Imai and Ratkovic, 2013). In this review, we focus on the intersection of this literature and OCEs. Assume each unit $i$ has a pair of potential outcomes $\big\{Y_i(1), Y_i(0)\big\}$ and a vector of pre-treatment covariates $X_i$, with $e(x) = Pr(W_i = 1|X_i = x)$ being the probability that a user is treated given a particular value of the covariate. For randomized studies where causality may be inferred, $e(X_i)$ is known and technically independent of $X_i$; however, when HTE analysis is under an observational setting, $e(X_i)$ is typically unknown. Most of the literature regarding HTEs employs the following assumptions: (1) SUTVA and (2) *unconfoundedness*, meaning that the response is independent of the treatment assignment $W_i$ conditional on the covariate, $\big\{Y_i(1), Y_i(0)\big\} \perp\!\!\!\perp W_i|X_i$. The main goal is to estimate the *conditional average treatment effect* (CATE), $\tau(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i(1)-Y_i(0)|X_i = x]$. Another key challenge is to detect exactly for which specific levels of the covariate $\tau(x)$ differs from $\tau$ and, given several covariates, identifying which $X$'s are the source of heterogeneity.

Interpretation is crucial in the online industry, thus a popular approach is to assume a linear mapping from $Y_i$ to $\big(W_i, X_i\big)$ from which main and interaction effects may be estimated. Unsurprisingly, the relationship between $Y_i$ and $X_i$ is often highly complex, thus a common method is to use the semi-parametric model from Robinson (1988), $Y_i = \tau(X_i)W_i+g(X_i)+\varepsilon_i$, which makes no assumptions about the forms of $\tau(X_i)$ and $g(X_i)$. Under unconfoundedness, one may write $Y_i - m(X_i) = \tau(X_i)(W_i - e(X_i)) + \varepsilon_i$, where $m(X_i) = \mathbb{E}[Y_i|X_i]$ and $e(X_i)$

are unknown. The $\ell_2$ loss function used to estimate heterogeneous treatment effects is $\hat{\tau}(X) = \text{argmin}_{\tau'} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[ Y_i - m(X_i) - \tau'(X_i)\big(W_i - e(X_i)\big) \right]^2 \right\}$. Thus HTE estimation is a ripe target for machine learning methods. Researchers have approached this problem using a technique called "Double Machine Learning" (DML) (Chernozhukov et al., 2017). Briefly, this technique models $m(X)$ and $e(X)$ as nuisance parameters, estimating them with nonparametric regression on a hold-out sample set. The CATE may then be estimated on the remaining sample set using a variety of machine learning methodologies. Chernozhukov et al. (2017) demonstrate that the above squared error loss is Neyman orthogonal to $m(X)$ and $e(X)$, which, along with sample splitting, ensures unbiasedness of $\hat{\tau}(X)$ and enforces parsimonious modeling of $Y$ with respect to the nuisance parameters. Syrgkanis et al. (2019) extend DML for estimating heterogeneous treatment effects when the covariates are hidden. Such situations arise in online experiments when users choose not to comply with a treatment (e.g., not clicking on a suggested article) due to unobserved factors. By modeling $Y$ with instrumental variables, Syrgkanis et al. (2019) estimate the HTEs using a doubly-robust, fully convex loss function that is minimized with an algorithm that builds on the DML technique. To avoid the challenge of directly estimating the HTE, Peysakhovich and Lada (2016) utilize historic, user-level data to learn individual effect estimates conditional on the covariates that correlate with the true treatment effect. Practitioners at Netflix also used DML to understand the localized impact on viewership from subbed and dubbed movies (Lan et al., 2022).

Other popular machine learning approaches for estimating $\tau(X)$ are regression trees and random forests. Following the DML approach, for instance, one may use trees to identify meaningful segments of a continuous or categorical variable, and then model $\tau(X)$ with partially linear regression. In an adaptation of the classical CART algorithm, Athey and Imbens (2016) build modified regression trees to partition the data into subgroups corresponding to different magnitudes of the treatment effect, thus each terminal leaf produces an estimate for $\tau(x)$, rather than the traditional estimate of $\mathbb{E}[Y|X_i = x]$. To correct for over-fitting,

an additional split of the training data into non-overlapping sub-partitions for each leaf is used. Naturally, this method can be extended to random forests to create a causal forest for estimating the HTE (Wager and Athey, 2018). While casual trees and forests do not require linearity of the treatment effects, and perhaps are conceptually more intuitive than DML, they are somewhat lacking in terms of interpretability compared to the effect estimates from DML and other similar methods. A further disadvantage is that the additional training split reduces the sample size for an application that may already suffer low power.

In addition to estimating $\tau(X)$, identifying which covariates or levels of covariates contribute to treatment heterogeneity is of great practical concern. Obtaining a parsimonious model of $Y$ is critical in such situations, as there are typically a large number of covariates from which to choose and strong statistical significance is required for detecting HTEs. This challenge is simultaneously a variable selection and multiple testing problem. Xie et al. (2018) assume an experimental design setup where $e(X) = Pr(W_i = 1)$, using this value to transform $Y_i$ into $Y_i^*$ such that $\mathbb{E}[Y_i^*|X] = \tau(X)$, which is estimated with the standard difference-of-means $\hat{\tau}$. Using $Y^* - \hat{\tau}$ as the response variable, the authors perform Lasso regression in conjunction with the "knockoff" variable selection defined by Barber, Candès, et al. (2015) to select heterogeneous covariates while controlling false discovery rates. They also demonstrate how to use the Benjamin-Hochberg correction to identify levels within these covariates where HTEs occur. Deng et al. (2016b) also use variable selection when clusters of covariates are of interest, such as device grouped by brand name. They employ a linear model with first order effect and second order interaction terms and enforce sparsity using total variation regularization, a technique similar to Fused Lasso (Petersen et al., 2016).

Given the wide array of scenarios under which HTEs occur in online experiments, there are still many situations where the methods discussed above may not be appropriate. Much of the literature in this review make strong model assumptions that are difficult to verify in practice. Additionally, the low power due to small effect sizes make multiple testing quite challenging. Simulations regarding the approach for controlling FDR in Xie et al.

(2018) showed that the knockoff method may be too conservative when faced with small effect sizes, and Deng et al. (2016b) reported difficulties regarding high false positive rates. For more open challenges regarding HTE estimation, we encourage the interested reader to consult Gupta et al. (2019), Kohavi et al. (2020), and Bojinov and Gupta (2022).

# 4 Long-Term Effects

**Motivating Example:** *At Bing, researchers hypothesised that generating large numbers of advertisements should have a positive effect on revenue, but may hurt user engagement in the long-term. To test this, the researchers exposed users to varying ad loads, noting a significant difference in engagement metrics for users exposed to a high ad load versus a low one. It was proposed that one may estimate the long-term effect by performing a post-hoc analysis some time after the experiment. Unfortunately, the post-hoc differences between high-load and low-load users could not be solely attributed to treatment assignment – many users quickly abandoned Bing as a result of too many ads, biasing results towards the users who remained (Dmitriev et al., 2016).*

Practitioners are often interested in understanding the treatment effect not just during the experiment, but months, even years after the experiment concludes. In many online experiments, the short-term treatment effect observed during and immediately after the experiment is not necessarily the same as the long-term effect. For instance, click-bait advertising has a positive short-term effect on click-through-rates, but a negative long-term effect on user retention and revenue (Kohavi et al., 2012). More generally, novelty and primacy effects are of concern. A novelty effect exists when a novel change is initially intriguing, leading to increased engagement, but that diminishes over time. A primacy effect on the other hand exists when the initial reaction to a change is not positive, but over time as users get used to the change their engagement increases (McFarland, 2012; Sadeghi et al., 2022). In both cases, the nature of the treatment effect may change over time as users learn.

Current OCE literature regarding long-term effect estimation is highly context-specific. At the time of writing this review, it is difficult to pinpoint a single statistical lineage of methodologies for this area (unlike with heterogeneous treatment effects, for example). The following section begins by introducing several distinct approaches that draw from a variety of statistical fields, and finishes with discussion of one area in particular that shows promise in providing a statistical framework for modeling and estimating long-term effects in online settings. For more industry-specific examples of the challenges concerning long-term effects, see Gupta et al. (2019) and Bojinov and Gupta (2022).

## 4.1 Review

A straightforward way to assess long-term effects is to simply run the experiment longer and ensure that the appropriate metrics for capturing long-term behavior are observed. However, much of the literature written by practitioners of OCEs has been devoted to describing the pitfalls associated with running long-term controlled experiments specifically for estimating long-term effects (Kohavi et al., 2009; Kohavi et al., 2012; Dmitriev et al., 2016; Gupta et al., 2019; Kohavi et al., 2020). Besides increased cost, several other external factors often make long-term experiments unappealing. For instance, when browser cookies are used to identify users, long-term experiments risk losing upwards of 75% of users as a result of cookie churn and are rendered invalid as a result (Dmitriev et al., 2016). These users may also re-enter the experiment unbeknownst to the experimenters and receive both the treatment and control experiences. This type of contamination can also happen if users access the product or service on multiple devices, a problem that becomes more likely as the experiment's duration increases. Additionally, the longer the experiment the more likely it is that multiple users (e.g., family members) will use the same device, obfuscating results. As such, in this section, we focus on techniques for estimating long-term treatment effects alternative to increasing experiment length.

Several approaches for estimating long-term effects intersect with other areas discussed in

this review. In Wang et al. (2019), long-term effects are characterized as a form of bias due to heterogeneous treatment effects (Section 3). In this context, long-term effects manifest because heavy-users (frequent users of the product) tend to be included in experiments at higher rates than light-users, biasing the ATE particularly in the short-term. Here, the treatment effect is presumed to be different depending on whether user $i$ is a heavy- or light-user. Under SUTVA and an assumed independence of outcomes from treatment assignment, the authors derive bias due to heavy-users in closed form, proposing a bias-adjusted jackknife estimator for the overall ATE. For a two-sided market where each experimental unit has a treatment history up to time $t$, Shi et al. (2020) leverage sequential testing (Section 5) and reinforcement learning to test for long-term treatment effects. Using data from a ride-sharing company, they demonstrate how their derived test statistic is able to detect long-term effects where regular two-sample t-tests fail. While the solutions from Wang et al. (2019) and Shi et al. (2020) are effective, they only target specific types of long-term effects, which limits their potential generalizability to other settings.

Another common solution is to define and measure short-term *driver metrics* that are causally linked to the long-term effect (Kohavi et al., 2020). Driver metrics allow practitioners to focus experiments on short-term goals while still taking into account the long-term effects (see Biddle (2019) for anecdotal examples). In Hassan et al. (2013), the authors define heuristics for modeling implicit indicators of customer satisfaction, noting that using query-based models instead of click-based models tend to serve as better proxies. Hohnhold et al. (2015) define models for how users "learn" to search or click for a product as a result of being exposed to a treatment, such as change in number of ads shown, using "learned click-through-rates" as a driver metric for estimating long-term effect on revenue. To estimate the effect on long-term revenue using short-term effects due to treatment, the authors model this as a linear function of short-term revenue and the estimated learned click-through-rates. The model has been successfully deployed by Google and is widely cited in the OCE literature (Kohavi et al., 2020; Gupta et al., 2019; Deng et al., 2017; Wang et al.,

2019). A recent paper by Sadeghi et al. (2022) proposes an observational approach based on difference-in-differences to estimate user learning and hence the long-term treatment effect in contexts where novelty and primacy effects exist.

Methodology in this area tends to resemble recent works in the causal inference literature that also aim to address this challenge by combining short-term experimental data with long-term observational data. This literature generally begins with the following. Assume a potential outcomes setup with two samples, $n_E$ (experimental) and $n_O$ (observational), with binary indicator $G_i \in \{E, O\}$. The tuple $(W_i,\ S_i,\ X_i)$ is observed in the experimental group and $(Y_i,\ S_i,\ X_i)$ in the observational group, where $S_i$ is an intermediate short-term outcome and $X_i$ is a pre-treatment covariate ($W_i$ may also be included in the observational group, see Athey et al. (2020a) and Imbens et al. (2022)). The goal is to estimate the average treatment effect of $W_i$ on $Y_i$, which is nontrivial since $Y_i$ is not observed in the experimental sample. The origins of this framework can be traced back to statistical literature regarding *surrogate outcomes*, used largely in biostatistics and econometrics (Prentice, 1989; Begg and Leung, 2000; Frangakis and Rubin, 2002; Ensor et al., 2016). The work by Athey et al. (2019) is one of the first papers that uses this framework for long-term effect estimation cited within the OCE community. The authors derive estimators of $\tau$ using $S_i$ as driver metrics and assume $W_i$ is not observable in $O$. They employ the "surrogate criterion", which requires that $Y_i$ be independent of $W_i$ given the short-term outcomes. It is straightforward to see that the approach in Hohnhold et al. (2015) is a special case of this approach, where $S_i$ is comprised of the learned click-through-rates and short-term revenue, $Y_i$ is long-term revenue, and the necessary conditions for estimating $\tau$ are unverified but implicitly assumed.

In practice, the surrogate criterion is notoriously tricky to satisfy. Athey et al. (2020b) relax this assumption by only requiring that $Y_i$ is independent of $W_i$ conditional on a *set* of surrogates, rather than on each individual surrogate. In perhaps one of the earliest publications using statistical surrogacy to estimate long-term effects specifically in OCEs, Cheng et al. (2020) show that one can relax the surrogacy assumption by extending this

framework to incorporate sequential testing. There is also evidence that some tech companies such as Facebook have used statistical surrogacy (Gupta et al., 2019), although it appears that too many surrogates may severely hamper interpretability. Recent work has shifted away from the surrogate criterion. Athey et al. (2020a) let $W_i$ be seen in the observational sample and estimate the treatment effect on $S_i$ in both samples, using the difference to adjust the ATE estimates. Imbens et al. (2022) consider a similar context and demonstrate how to account for unmeasured confounding variables that impact treatment, short-term, and long-term outcomes. Further exploration of combining short-term experimental data with observational data to estimate long-term effects may show promise with respect to OCE applications.

# 5 Optional Stopping

**Motivating Example:** *Suppose an online streaming service is altering a certain feature that positively correlates with subscription renewals. While an improvement to this feature could increase the rate of subscription renewals, a harmful change may have the opposite effect. It is in the service's best interest to quickly abandon harmful or poorly performing variants, and identify those that perform well. Methods that support early termination without compromising overall statistical validity are desirable. A notable practice within this context is to "ramp up" the experiment by slowly exposing an increasing percentage of users to the treatment (Xu et al., 2018).*

Most OCEs are run in real-time, and it is not uncommon for estimates and confidence intervals associated with $\tau$, and p-values associated with $H_0 : \tau = 0$, to be updated in near-real-time as the data is collected. Although a fixed horizon is typically determined based on development cycles (typically two weeks) and minimal sample size requirements (determined via power arguments), the near-real-time availability of results encourages a phenomenon known colloquially as "peeking", whereby p-values are monitored continuously

and the experiment is stopped as soon as a significant p-value is observed. While it is well known that this practice seriously inflates false positive rates (Johari et al., 2017; Kohavi et al., 2022), there are nevertheless situations where having a mechanism for *optional stopping* is desirable. For example, it is extremely important to quickly detect and abort treatments that are negatively impacting the user experience (Lindon et al., 2022). Thus, in situations like this, a methodology that permits near-real-time decision-making without inflating Type I error rates is invaluable.

Recently there has been increased emphasis on *sequential testing* methods which assess $H_0 : \tau = 0$ using sample size-dependent decision rules. Within this class of methods, Type I error is controlled at each current sample size $n$, which avoids the inflated risk of Type I error that is associated with preemptively stopping an experiment when the current p-value is statistically significant by chance. Such methods improve testing efficiency due to the lower sample size a sequential test will terminate at, on average, regardless of where the true treatment effect might lie. However, there is no free lunch. Existing methodology is not well-suited for all OCE applications, such as monitoring multiple metrics (e.g., the OEC and guardrails). Additionally, the reduced sample sizes guarantee under-powered HTE inference across user segments. Nevertheless, sequential testing methods have appeared in numerous OCE applications and studies (Kohavi et al., 2013; Johari et al., 2017; Kharitonov et al., 2015; Deng et al., 2016a; Yu et al., 2020; Shi et al., 2020; Abhishek and Mannor, 2017; Ju et al., 2019). The following section broadly introduces the method of sequential testing as it pertains to ongoing evaluation of the treatment effects(s) of interest in OCEs.

## 5.1 Review

The vast majority of the OCE literature in sequential testing builds on the classic *sequential probability ratio test* (SPRT) developed by Wald (1945). Define constants $0 < B < A$ where $B = \frac{\beta}{1-\alpha}$ and $A = \frac{1-\beta}{\alpha}$, and a simple hypothesis test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. The SPRT method proceeds as follows. For current sample size $n$, compute the likelihood ratio

test statistic $\Lambda_n = \prod_{i=1}^{n} \frac{f(y_i|\theta_1)}{f(y_i|\theta_0)}$, where $y_i$ are observations of i.i.d data $\{Y_i\}_{i=1}^{n} \sim f(\cdot|\theta)$. The rejection region divides the sample space into three mutually exclusive decision rules: (1) if $\Lambda_n > A$, reject $H_0$ and stop the test. (2) If $\Lambda_n < B$, fail to reject $H_0$ and stop the test. (3) If $B < \Lambda_n < A$, obtain another observation $Y_{n+1}$ and compute $\Lambda_{n+1}$. Although it seems like testing in this manner would permit the possibility of never drawing a conclusion about $H_0$ (i.e., $n \to \infty$), Wald (1947) proved that the SPRT will eventually terminate for finite $n$. SPRT does not require specifying $n$ in advance, and requires on average about half the number of observations required for a uniformly most powerful Neyman-Pearson test for the same level of power (Wald, 1945).

The first and perhaps most widely-known application of sequential testing in OCEs is a modified version of SPRT called the *mixture sequential probability ratio test*, or mSPRT (Johari et al., 2017; Pramanik et al., 2021). The mSPRT allows for a simple null hypothesis versus a composite alternative hypothesis $H_1 : \theta \neq \theta_0$ by assuming a mixture distribution $H$ with density $h(\cdot)$ defined over the parameter space of all possible $\theta$. The test statistic is therefore a mixture of the likelihood ratios, $\Lambda_n = \int \prod_{i=1}^{n} \frac{f(y_i|\theta)}{f(y_i|\theta_0)} h(\theta) d\theta$. The procedure rejects $H_0$ and ends if $\Lambda_n^H \geq \alpha^{-1}$. Johari et al. (2017) use the mSPRT to define "always valid p-values", which are computed iteratively such that $p_0 = 1$; $p_n = min\{p_{n-1}, (\Lambda_n^H)^{-1}\}$. Thus, practitioners may stop an experiment at any time while still controlling Type I error. The always valid p-values and their confidence interval counterparts are currently deployed by Optimizely, a widely-used third-party vendor for OCEs (Pekelis, 2015). With the vast quantity of experiments that this company has facilitated, they are able to leverage prior data for estimating the mixture distribution $H$. However, Johari et al. (2019) derive their optimality conditions for mSPRT only for data that comes from the exponential family of distributions, which does not cover distributions for ratio metrics, which are popular in industry. Another major limitation of these always valid p-values lies in how they define the likelihood ratios for a two-sample hypothesis test. While the authors assume a standard two-sample stream of data, they impose an additional restriction by arbitrarily pairing observations in accor-

dance with a matched pairs design. While this allows for defining a tractable $f(y_i|\theta)$, there is arguably very little reason to apply a matched pairs design to a two-sample hypothesis test. Thus far, discussion of this particular issue has not been prevalent in the relevant literature.

The well-publicized usage of mSPRT has inspired several related works in the literature. Abhishek and Mannor (2017) account for the situation where $f(\cdot|\theta)$ is unknown by creating a bootstrap algorithm to approximate $\Lambda_n^H$. While the algorithm also requires a prior distribution to approximate $H$, this method still allows practitioners to use mSPRT for commonly used online metrics that are otherwise difficult to model. The work in Lindon and Malek (2020) extends mSPRT to multinomial count data, which includes an application for conducting SRM tests sequentially, in near-real-time. Yu et al. (2020) also extend mSPRT to the multiple testing scenario to test for heterogeneous treatment effects, using always valid p-values to allow for continuous monitoring. In a well-received paper, Xu et al. (2018) use a technique similar to mSPRT, called a *generalized sequential probability ratio test* (GSPRT), to determine the risk of exposing more users to a new variant. Briefly, the GSPRT uses the supremums of the likelihoods in $\Lambda_n$ and can be shown to require smaller sample sizes on average than mSPRT (Chan and Lai, 2005). Xu et al. (2018) use a prior-weighted GSPRT to provide a rigorous statistical framework dubbed "speed, quality, and risk" (SQR) for the practice of ramping up, gradually introducing users to a new variant in order to mitigate the fallout associated with exposing them to potentially negative variants (for a high-level discussion of SQR, see Chapter 15 of Kohavi et al. (2020)). An alternative to frequentist sequential testing is also explored by Deng et al. (2016a), where the authors use Bayesian hypothesis testing as the foundation. Bayesian methods for OCEs are briefly discussed in Section 7. Finally, we acknowledge that the sequential testing methods discussed here are fully sequential. See Georgiev (2022) and the references contained therein for a discussion of the value of group sequential tests as an alternative framework.

# 6 Network Interference

**Motivating Example:** *Suppose LinkedIn plans to test the impact of a new feature for their messaging service, with the objective to increase total messages sent. Using balanced randomization, given that user $i$ is exposed to the new feature, there is approximately a 50% chance said user's friend $j$ is randomized to the old service. Under this scenario and if the new feature indeed increases messages sent, it is likely that friend $j$ will also send more messages in response to $i$, despite $j$ belonging to the old service. Thus, the overall impact of the new messaging feature on total messages sent is confounded by the network interference between treatment and control groups, biasing standard estimators for the ATE (Saint-Jacques, 2019).*

Recall that SUTVA requires that the potential outcome $Y_i(W_i)$ for unit $i$ remain the same regardless of the treatment assignments and outcomes of the other experimental units. However, in certain OCE applications (e.g., social networks and two-sided marketplaces), SUTVA may be violated (Gui et al., 2015; Chamandy, 2016; Spang et al., 2021). In the specific case where the units are connected to one another through a *network*, SUTVA violations are referred to as either *network exposure*, *network effects*, *network interference* or *spillover effects*. Kohavi et al. (2020) refer to this type of interference as being a result of "direct" connections; they contrast this with interference (and SUTVA violations) due to "indirect" connections that may arise in shared resource problems such as those evident in online marketplaces, auctions, and ad campaigns (Blake and Coey, 2014; Holtz et al., 2020; Liu et al., 2021). Our focus here is on network interference, so we refer the reader to Chapter 22 of Kohavi et al. (2020) and the references contained therein for a deeper discussion of indirect interference.

OCEs where the experimental units are subject to network exposure are known as *network A/B tests*, where users and the connections among them are modeled by a network $\mathcal{G}$, with $n \times n$ adjacency matrix $\mathbf{A} = [A_{ij}]$. In most OCE settings, $\mathbf{A}$ is assumed to be fixed and observable, although situations where this is not the case are also considered (Egami, 2017).

The goal of estimating the ATE remains of primary interest. However, when SUTVA is violated, standard randomization schemes and estimators tend to ignore the network effect, which typically produces biased estimates of $\tau$. Consider the following example: suppose the response $Y_i = \alpha + \beta W_i + \gamma S_i + \varepsilon_i$ is linearly related to the treatment effect $\beta$ and network spillover effect $\gamma$, where $S_i$ is the proportion of $i$'s neighbors that received treatment. The ATE is therefore $\beta + \gamma$, since $\mathbb{E}[S_i|W_i = 1] = 1$ and $\mathbb{E}[S_i|W_i = 0] = 0$. Under the usual balanced randomization, however, $\mathbb{E}[S_i] = 0.5$ for both treatment and control groups, thus the expected value of the usual difference of means estimator $\hat{\tau}$ is $\beta$, which has a bias of $\gamma$. Generally, the exact form of the ATE depends on the assumed structure of $\mathcal{G}$ and definition of $S_i$; similarly for the form and bias of $\hat{\tau}$. Thus, there are two major problems in network A/B testing that current research aims to address: (1) modeling and estimating the network spillover effect, and (2) optimal treatment allocation for producing unbiased estimates of $\tau$ in the presence of network interference. Reviewing work in these areas is the focus of the following subsection

## 6.1 Review

A commonly proposed approach for dealing with network effects in OCEs is to randomize treatments with *graph cluster randomization* (Karrer et al., 2021; Eckles et al., 2014; Gui et al., 2015; Saveski et al., 2017; Sangho Yoon, 2018; Zhou et al., 2020; Ugander et al., 2013). With cluster-based randomization, the network is partitioned into subgroups or *clusters*, such that edge connectivity within clusters is higher than between clusters. Network partitioning, also known as community detection in network science, is a well-researched area, with most OCEs using established graph clustering algorithms as found in Newman (2006), Leskovec et al. (2010), and Mucha et al. (2010) and Stanley et al. (2016). Treatments are then randomized to users at the cluster level with the standard difference of means estimator, a common choice for estimating the ATE. Eckles et al. (2014) explore several linear models for relating user response to the network effect, and perform a suite of simulations that show graph cluster

randomization reduces bias when compared to naive random allocation. They also provide a theorem that shows the bias from network effects will always be less than or equal to the bias from random allocation, assuming $Y_i = \alpha + \beta W_i + \gamma S_i + \varepsilon_i$. Gui et al. (2015) draw from this work, modeling the response as $Y_i = \alpha + \beta W_i + \gamma \sum_{j=1}^{n} A_{ij} W_j + \eta \sum_{j=1}^{n} A_{ij} Y_j / d_i$, where $d_i$ is the degree of node $i$, $\gamma$ is the spillover effect, and $\eta$ describes how users tend to exhibit behavior similar to their neighbors'. They showed that with a network sampled such that clusters are "balanced", where clusters are all equal in size, one can eliminate the bias in $\hat{\tau}$. Their new algorithm for balanced cluster-based randomization was empirically shown to reduce bias, although theoretical justification was not provided. To address the question of how to detect when the spillover effect is present, Saveski et al. (2017) develop a model-free two stage cluster-randomization design for testing for the presence of SUTVA violations, and Athey et al. (2018) derive exact p-values for nonsharp null hypotheses of no spillover effects. Recent work by Karrer et al. (2021) utilizes imbalanced clusters with a regression-adjusted estimator, along with a post-analysis framework that is also used to detect network effects.

While Gui et al. (2015) use a common framework for OCE applications, the linear model assumption is known to be quite restrictive, particularly for network applications. Basse and Airoldi (2018) specifically study the drawbacks of traditional parametric assumptions for modeling network effects. Some practioners instead use *network exposure models* to model the spillover effect (Backstrom and Kleinberg, 2011; Katzir et al., 2012). Network exposure models define a set of conditions for each $i$ under which the spillover effects from $i$'s neighbors are the same. For example, the *neighborhood exposure model* from Backstrom and Kleinberg (2011) and Gui et al. (2015) estimates $\tau$ with $\frac{1}{|N_1^\theta|} \sum_{i \in N_1^\theta} Y_i - \frac{1}{|N_0^\theta|} \sum_{i \in N_0^\theta} Y_i$, where $\sigma_i$ is the percent of neighbors of $i$ that received treatment, $N_1 = \{i : W_i = 1, \sigma_i \geq \theta\}$, $N_0 = \{i : W_i = 0, \sigma_i \leq 1 - \theta\}$, and $\theta \in [0, 1]$. With network exposure models, one need not make explicit assumptions about how the spillover effect relates to the response, although the corresponding ATE estimators tend to be more complex. Ugander et al. (2013) catalogue the various network exposure models that have been commonly adopted in the literature

(Eckles et al., 2014; Gui et al., 2015; Saveski et al., 2017).

While cluster-based randomization approaches are commonly used in practice, the limitations of this method are significant enough that researchers remain interested in alternative approaches. First, because this approach uses clusters as the experimental units and cluster counts typically are far smaller than the total number of users, experiments under this approach tend to lack adequate power. To mitigate this, Saint-Jacques et al. (2019) propose sampling many "ego-networks", which are *small* clusters comprised of a central user and a carefully selected subset of their immediate neighbors. Second, the majority of online social networks are highly dense, making it extremely difficult to obtain reasonably isolated clusters that are representative of the true network. Nandy et al. (2020) avoid explicit model assumptions by defining $\mathcal{G}$ as a directed network of producers $j$ and consumers $i$. Treatment intervention $(r)$ is represented by rewiring edge probabilities by replacing the original $p_{ij}^{base}$ with $p_{ij}^{(r)}$, where $p_{ij} = Pr(A_{ij} = 1)$. Nandy et al. (2020) use this setup to frame treatment allocation as an optimization problem, where treatments are randomized such that the effect from network exposure under the new treatment is as small as possible. Their method showed an improvement over cluster-based randomization in terms of bias of the ATE, particularly for highly dense networks. Note Nandy et al. (2020) and Saint-Jacques et al. (2019) and Gui et al. (2015) all assume that the network is known, where in fact it is highly possible there are unobserved covariates or network effects influencing network structure and user response. Bajari et al. (2021) employ the producer-consumer marketplace set-up to address interference without a network model. Rather, users are assumed to belong to a number of different populations that serve as indices for the outcomes and treatment assignments. Bajari et al. (2021) define a new class of experimental designs, *Multiple Randomization Designs*, that model the response as a tuple with elements corresponding to each population and randomize treatments at the tuple-level.

Despite the drawbacks of defining a parametric model for $Y_i$, there are inherit advantages to this approach, such as analyzing heterogeneity in the form of interactions or applying

conventional tools like censoring and stratification (Walker and Muchnik, 2014). Under this framework, a natural solution to the question of treatment allocation is optimal design of experiments (DoE) theory. Optimal DoE refers to the general practice of choosing a design matrix from the space of potential candidates, $X \in \mathcal{X}$, according to various optimality criterion. In Parker et al. (2017), the response is modelled as $Y_i = \alpha + \tau_{t(i)} + \sum_{j=1}^{n} A_{ij}\gamma_{t(j)} + \varepsilon_i$, where $\tau_{t(i)}$ represents the treatment applied to $i$, assuming $t \in \{1, ...T\}$ treatments. A blocking parameter $b_i$ can also be introduced to this model (Koutra, 2017). With this framework, Parker et al. (2017) and Koutra (2017) provide some interesting insights into what optimal designs for network A/B testing might look like, namely that unbalanced designs tend to be better at reducing the variance of $\hat{\tau}_j$ than balanced allocation. However, these models are rather unrealistic. Because they do not scale the spillover effect by the degree of node $i$, as the number of neighbors of $i$ grows, $\sum_{j=1}^{n} A_{ij}\gamma_{t(j)} \to \infty$ as well, meaning the spillover effect completely dominates $\tau_{t(i)}$ for the large networks typically observed in OCEs. Parker et al. (2017) and Koutra (2017) also do not optimize for the ATE, instead considering optimal designs for only $\tau_j$ by minimizing the average variance of all pairwise treatment effects. Additionally, these optimal designs are chosen with an exhaustive search algorithm, which searches the entire space of $\mathcal{X}$, or $T^n$ potential designs, before selecting $X$. Indeed, some of the designs obtained via search algorithm in Parker et al. (2017) were outperformed by randomly generating $X$. Pokhiko et al. (2019) and Zhang and Kang (2020) alternatively choose conditional auto-regressive models to mimic the network effect by correlating the response error of $i$ with that of its neighbors. A strong limitation of this approach is this correlation is assumed to be the same across all nodes. Zhang and Kang (2020) address this issue by using Bayesian priors via simulation, but do not leverage network information in defining them. Given that optimal design of experiments is still relatively new to network A/B testing and OCEs in general, we believe that there may be many opportunities for DoE methods in this area.

# 7 Beyond This Review

We have presented literature that generally assumes a single treatment and control under a frequentist framework. While this setting describes an appreciable majority of OCEs, there is also growing interest in methodologies that extend beyond the scope of this review. Researchers aiming to circumvent limitations of the frequentist p-value have turned to Bayesian methods (Stucchio, 2015; Letham et al., 2019; Deng et al., 2016a; Deng et al., 2021a; Kamalbasha and Eugster, 2021; Hoffmann and Wagenmakers, 2021), including implementations of Bayes factor hypothesis testing (Deng, 2015) and tests for practical significance (Stevens and Hagar, 2022). Many practitioners have noted that the ATE itself is not a quantity of interest in several applications, e.g., when optimizing tail performance, and have begun to develop approaches using *quantile metrics* (Liu et al., 2019; Howard and Ramdas, 2019; Lux, 2018). *Multi-armed bandits* have been used to handle multiple treatments in online settings, with a focus on sequential decision-making and exposing more users to successful variants to increase reward (Liu et al., 2014; Issa Mattos et al., 2019; Birkett, 2019; Amadio, 2020; Lomas et al., 2016). Thompson sampling (Scott, 2010; Scott, 2015; Dimakopoulou et al., 2021) as well as contextual bandits (Li et al., 2010; Agarwal et al., 2016) have all been used in industry.

Although OCEs with multiple variants are reasonably common, full-factorial experiments that emphasize estimation of main and interaction effects are uncommon; Kohavi et al. (2009) and Georgiev (2019) argue that the added practical complexity of a full-factorial experiment is not worth the potential insights gained when comparable insights can be had by running multiple single-factor experiments concurrently. Nevertheless, *multivariate tests* (where the multiple variants are defined by the factorial enumeration of multiple factors' levels) *do* exist in this space (McFarland, 2012; Wildman, 2019), but the goal of the analysis is primarily to identify the optimal variant, *not* to estimate individual effects. Though multivariate tests are not as common as A/B or A/B/n tests, research in this area carries on (Sadeghi et al., 2019), with recent research in optimal design (Bhat et al., 2020), nonparametric estimators

for panel experiments (Bojinov et al., 2021), and factorial designs for sequential testing (Haizler and Steinberg, 2020). How to avoid, identify, and estimate interactions between multiple concurrent experiments is also of great interest (Kohavi et al., 2009; Gupta et al., 2019; Chan, 2021).

Another important facet of OCEs outside the scope of this review is the issue of ethics (Gupta et al., 2019). As noted, the experimental units in OCEs are often people – human subjects – and so a salient concern is whether experiments involving them are ethical. Many OCEs test harmless interface changes, but there exist A/B tests that through *code* induce *deception*, thus named C/D tests (Benbunan-Fich, 2017; Kontotasiou, 2021). One example is Facebook's infamous *emotional contagion* experiment in which the sentiment of content shown in nearly 700,000 users' News Feeds was altered to determine whether this impacted their own emotions (Kramer et al., 2014). Another example is OKCupid's *power of suggestion* experiment in which matched users were told their compatibility was higher than what the matching algorithm predicted in order to investigate the impact of simply telling couples they're a good match (Rudder, 2014). A more recent example is LinkedIn's *strength of weak ties* experiment in which the "People You May Know" algorithm for 20 million users was modified to intentionally vary the quality of the recommendations, potentially reducing job opportunities and job mobility for some users (Rajkumar et al., 2022). The primary concern in these settings is informed consent; users generally do not know when they're being experimented on, nor do they necessarily have a way to opt out of such an experiment. They implicitly consent to such experimentation when they agree to a service's terms and conditions, however, whether such consent is *informed* is debatable (Benbunan-Fich, 2017). Academics involved in human subjects research will be familiar with institutional review boards (IRBs) and ethics clearance. Such formal oversight is generally absent in the private sector. However, Kohavi et al. (2020) do advocate for the establishment of processes that fulfill this purpose so that an experiment's risks and benefits are carefully considered, and transparent protocols for informed consent and drop-out are instated. The authors also

advocate for tools, infrastructure, and processes to ensure data security and data privacy, another issue especially relevant in this day and age. See Kohavi et al. (2020) and Bojinov and Gupta (2022) for expanded discussions of identified data, anonymous data, re-identification, and differential privacy in the context of OCEs.

# 8 Conclusion

We conclude this literature review with a call to action for greater collaboration between industry and academic statisticians to address the research challenges presented by online experimentation. While this paper may be one of the first to provide a cohesive review of the OCE statistics literature, the need for increased cooperation between industry and academia has already been explicitly stated by experts at thirteen leading organizations that run online experiments (Gupta et al., 2019). We hope this review contributes to this goal by introducing academicians to the context and goals of online experimentation, as well as providing examples and broad, technical discussion of the statistical methodologies regarding sensitivity, effect size, heterogeneity, long-term effects, optional stopping, and network interference.

The value of experimentation and the accompanying philosophy of trial and error has been observed in many facets of society (Manzi, 2012), and its positive impacts in the realm of business in particular, are remarkable (Koning et al., 2022). Online controlled experiments are vital tools utilized hundreds of times a day by companies whose products touch the lives of billions (Kohavi et al., 2013; Xu et al., 2015; Google, 2022). As many vital societal functions shift online at an unprecedented rate, online experimentation has already found applications outside the mainstream spheres of technology and e-commerce. OCEs have been used to optimize political advertisements and increase user engagement with campaign platforms during the Obama and Trump elections (Christian, 2012; Bump, 2019). Decision-making tools streamlined by OCEs help clinicians make safer, more cost

effective decisions regarding patient care (Austrian et al., 2021). OCEs have also been deployed to identify the psychological impacts of social media on younger demographics (Isaac, 2021). Along with the significant growth and popularity of careers under the evolving "data science" profession, online experimentation is almost certainly going to become a common tool for online businesses of all sizes (Schroeder, 2021). Given the breadth and depth of OCE applications, we believe that solving the research challenges presented in this review will improve the quality of data-driven decision making in online businesses across the applied domain.

# References


Abhishek and Mannor (2017). "A Nonparametric Sequential Test for Online Randomized Experiments". *Proceedings of the 26th International Conference on World Wide Web Companion*. WWW '17 Companion. Perth, Australia: International World Wide Web Conferences Steering Committee, pp. 610–616. ISBN: 9781450349147. DOI: 10.1145/3041021.3054196.

Acharya, Prakash, Saxena, and Nigam (2013). "Sampling: Why and how of it". *Indian Journal of Medical Specialties* 4.2, pp. 330–333.

Agarwal, Bird, Cozowicz, Hoang, Langford, Lee, Li, Melamed, Oshri, Ribas, et al. (2016). "Making contextual decisions with low technical debt". *arXiv preprint arXiv:1606.03966*.

Amadio (2020). *Multi-Armed Bandits and the Stitch Fix Experimentation Platform*. https://multithreade (Accessed on 03/21/2022).

Athey, Chetty, and Imbens (2020a). *Combining Experimental and Observational Data to Estimate Treatment Effects on Long Term Outcomes*. DOI: 10.48550/ARXIV.2006.09676.

Athey, Chetty, Imbens, and Kang (2020b). *Estimating Treatment Effects using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index*. arXiv: 1603.09326 [stat.ME].

Athey, Chetty, Imbens, and Kang (2019). *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely*. Tech. rep. National Bureau of Economic Research.

Athey, Eckles, and Imbens (2018). "Exact p-values for network interference". *Journal of the American Statistical Association* 113.521, pp. 230–240.

Athey and Imbens (2016). "Recursive partitioning for heterogeneous causal effects". *Proceedings of the National Academy of Sciences* 113.27, pp. 7353–7360. ISSN: 0027-8424. DOI: 10.1073/pnas.1510489113. eprint: https://www.pnas.org/content/113/27/7353.full.pdf.

Austrian, Mendoza, Szerencsy, Fenelon, Horwitz, Jones, Kuznetsova, and Mann (2021). "Applying A/B Testing to Clinical Decision Support: Rapid Randomized Controlled Trials". *Journal of medical Internet research* 23.4, e16651.

Backstrom and Kleinberg (2011). "Network bucket testing". *Proceedings of the 20th international conference on World wide web*. WWW '11. New York, NY, USA: Association for Computing Machinery, pp. 615–624. ISBN: 978-1-4503-0632-4. DOI: 10.1145/1963405.1963492.

Bajari, Burdick, Imbens, Masoero, McQueen, Richardson, and Rosen (2021). "Multiple Randomization Designs". *arXiv preprint arXiv:2112.13495*.

Barber, Candès, et al. (2015). "Controlling the false discovery rate via knockoffs". *The Annals of Statistics* 43.5, pp. 2055–2085.

Basse and Airoldi (2018). "Limitations of Design-based Causal Inference and A/B Testing under Arbitrary and Network Interference". *Sociological Methodology* 48.1. Publisher: SAGE Publications Inc, pp. 136–151. ISSN: 0081-1750. DOI: 10.1177/0081175018782569.

Begg and Leung (2000). "On the use of surrogate end points in randomized trials". *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 163.1, pp. 15–28.

Benbunan-Fich (2017). "The ethics of online research with unsuspecting users: From A/B testing to C/D experimentation". *Research Ethics* 13.3-4, pp. 200–218.

Berman and Van den Bulte (2021). "False discovery in A/B testing". *Management Science*.

Bhat, Farias, Moallemi, and Sinha (2020). "Near-Optimal A-B Testing". *Management Science*. Publisher: INFORMS. ISSN: 0025-1909. DOI: 10.1287/mnsc.2019.3424.

Biddle (2019). *Proxy Metrics: How to define a metric to prove or disprove your hypotheses and measure progress*. https://gibsonbiddle.medium.com/4-proxy-metrics-a82dd30ca810. (Accessed on 03/04/2022).

Birkett (2019). *When to Run Bandit Tests Instead of A/B/n Tests*. CXL. Library Catalog: cxl.com. URL: https://cxl.com/blog/bandit-tests/ (visited on 06/16/2020).

Blake and Coey (2014). "Why marketplace experimentation is harder than it seems: The role of test-control interference". *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 567–582.

Bojinov and Gupta (2022). "Online experimentation: Benefits, operational and methodological challenges, and scaling guide". *Harvard Data Science Review* 4.3.

Bojinov, Rambachan, and Shephard (2021). "Panel experiments and dynamic causal effects: A finite population perspective". *Quantitative Economics* 12.4, pp. 1171–1196.

Boucher, Knoblich, Miller, Patotski, and Saied (2020). *Google Analytics Solutions — Optimize*. https://www.microsoft.com/en-us/research/group/experimentation-platform-exp/arti (Accessed on 09/16/2022).

Box, Hunter, and Hunter (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*. 2nd ed. Wiley-Interscience.

Bump (2019). *Analysis — '60 Minutes' profiles the genius who won Trump's campaign: Facebook*.

Chamandy (2016). *Experimentation in a Ridesharing Marketplace*.

Chan and Lai (2005). "Importance Sampling for Generalized Likelihood Ratio Procedures in Sequential Analysis". *Sequential Analysis* 24.3, pp. 259–278. DOI: 10.1081/SQA-200063280. eprint: https://doi.org/10.1081/SQA-200063280.

Chan (2021). *Embrace Overlapping A/B Tests and Avoid the Dangers of Isolating Experiments*. https://blog.statsig.com/embracing-overlapping-a-b-tests-and-the-danger-of-isol (Accessed on 03/21/2022).

Chen, Liu, and Xu (2018). "Automatic Detection and Diagnosis of Biased Online Experiments". *arXiv preprint arXiv:1808.00114*.

Cheng, Guo, and Liu (2020). "Long-Term Effect Estimation with Surrogate Representation". *arXiv preprint arXiv:2008.08236*.

Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey (2017). "Double/debiased/neyman machine learning of treatment effects". *American Economic Review* 107.5, pp. 261–65.

Christian (2012). "The A/B Test: Inside the Technology That's Changing the Rules of Business". *Wired* 20.5. ISSN: 1059-1028.

Courthoud (2022). *Understanding CUPED*. https://towardsdatascience.com/understanding-cuped-a (Accessed on 08/18/2022).

Crook, Frasca, Kohavi, and Longbotham (2009). "Seven pitfalls to avoid when running controlled experiments on the web". *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*. the 15th ACM SIGKDD international conference. Paris, France: ACM Press, p. 1105. ISBN: 978-1-60558-495-9. DOI: 10.1145/1557019.1557139.

Deng (2015). "Objective Bayesian Two Sample Hypothesis Testing for Online Controlled Experiments". *Proceedings of the 24th International Conference on World Wide Web*. WWW '15 Companion. Florence, Italy: Association for Computing Machinery, pp. 923–928. ISBN: 978-1-4503-3473-0. DOI: 10.1145/2740908.2742563.

Deng and Hu (2015). "Diluted Treatment Effect Estimation for Trigger Analysis in Online Controlled Experiments". *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 349–358.

Deng, Li, Lu, and Ramamurthy (2021a). "On Post-selection Inference in A/B Testing". *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2743–2752.

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

Deng, Lu, and Litz (2017). "Trustworthy Analysis of Online A/B Tests: Pitfalls, challenges and solutions". *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. WSDM '17. Cambridge, United Kingdom: Association for Computing Machinery, pp. 641–649. ISBN: 978-1-4503-4675-7. DOI: 10.1145/3018661.3018677.

Deng and Shi (2016). "Data-Driven Metric Development for Online Controlled Experiments: Seven Lessons Learned". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM, pp. 77–86. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939700.

Deng, Xu, Kohavi, and Walker (2013). "Improving the sensitivity of online controlled experiments by utilizing pre-experiment data". *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*. the sixth ACM international conference. Rome, Italy: ACM Press, p. 123. ISBN: 978-1-4503-1869-3. DOI: 10.1145/2433396.2433413.

Deng, Yuan, and Salama-Manteau (2021b). *Variance Reduction for Experiments with One-Sided Triggering using CUPED*. arXiv: 2112.13299 [stat.ME].

Deng, Zhang, Chen, Kim, and Lu (2016b). "Concise summarization of heterogeneous treatment effect using total variation regularized regression". *arXiv preprint arXiv:1610.03917*.

Dimakopoulou, Ren, and Zhou (2021). "Online Multi-Armed Bandits with Adaptive Inference". *Advances in Neural Information Processing Systems* 34.

Dmitriev, Frasca, Gupta, Kohavi, and Vaz (2016). "Pitfalls of long-term online controlled experiments". *2016 IEEE international conference on big data (big data)*. IEEE, pp. 1367–1376.

Dmitriev, Gupta, Kim, and Vaz (2017). "A Dirty Dozen: Twelve Common Metric Interpretation Pitfalls in Online Controlled Experiments". *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '17. New York, NY, USA: Association for Computing Machinery, pp. 1427–1436. ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3098024.

Drutsa, Gusev, and Serdyukov (2015a). "Future user engagement prediction and its application to improve the sensitivity of online experiments". *Proceedings of the 24th International Conference on World Wide Web*, pp. 256–266.

Drutsa, Ufliand, and Gusev (2015b). "Practical Aspects of Sensitivity in Online Experimentation with User Engagement Metrics". *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. CIKM '15. Melbourne, Australia: Association for Computing Machinery, pp. 763–772. ISBN: 9781450337946. DOI: 10.1145/2806416.2806496.

Eckles, Karrer, and Ugander (2014). "Design and analysis of experiments in networks: Reducing bias from interference". *arXiv:1404.7530 [physics, stat]*. arXiv: 1404.7530.

Egami (2017). "Unbiased estimation and sensitivity analysis for network-specific spillover effects: Application to an online network experiment". *arXiv preprint arXiv:1708.08171*.

Ensor, Lee, Sudlow, and Weir (2016). "Statistical approaches for evaluating surrogate outcomes in clinical trials: a systematic review". *Journal of biopharmaceutical statistics* 26.5, pp. 859–879.

Fabijan, Dmitriev, Holmstrom Olsson, and Bosch (2018). "Online Controlled Experimentation at Scale: An Empirical Survey on the Current State of A/B Testing". *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pp. 68–72. DOI: 10.1109/SEAA.2018.00021.

Frangakis and Rubin (2002). "Principal stratification in causal inference". *Biometrics* 58.1, pp. 21–29.

Georgiev (2022). *Fully Sequential vs Group Sequential Test*. https://blog.analytics-toolkit.com/2022/

Georgiev (2019). *Statistical methods in online A/B testing*. Self-Published.

Google (2022). *How Google's Algorithm is Focused on Its Users - Google Search*. https://www.google.com/ (Accessed on 03/29/2022).

Gui, Xu, Bhasin, and Han (2015). "Network A/B Testing: From Sampling to Estimation". *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. Florence, Italy: International World Wide Web Conferences Steering Committee, pp. 399–409. ISBN: 978-1-4503-3469-3. DOI: 10.1145/2736277.2741081.

Gupta, Kohavi, Tang, Xu, Andersen, Bakshy, Cardin, Chandran, Chen, Coey, Curtis, Deng, Duan, Forbes, Frasca, Guy, Imbens, Saint Jacques, Kantawala, Katsev, Katzwer, Konutgan, Kunakova, Lee, Lee, Liu, McQueen, Najmi, Smith, Trehan, Vermeer, Walker, Wong, and Yashkov (2019). "Top Challenges from the First Practical Online Controlled Experiments Summit". *SIGKDD Explor. Newsl.* 21.1, pp. 20–35. ISSN: 1931-0145. DOI: 10.1145/3331651.3331655.

Haizler and Steinberg (2020). "Factorial Designs for Online Experiments". *Technometrics*, pp. 1–12. ISSN: 0040-1706, 1537-2723. DOI: 10.1080/00401706.2019.1701556.

Hassan, Shi, Craswell, and Ramsey (2013). "Beyond clicks: query reformulation as a predictor of search satisfaction". *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pp. 2019–2028.

Hern (2014). *Why Google has 200m reasons to put engineers over designers — Google — The Guardian*. https://www.theguardian.com/technology/2014/feb/05/why-google-engineers-desi (Accessed on 10/26/2021).

Hoffmann and Wagenmakers (2021). "Bayesian inference for the A/B test: Example applications with r and jasp". *PsyArXiv. June* 10.

Hohnhold, O'Brien, and Tang (2015). "Focusing on the Long-term: It's Good for Users and Business". *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. Sydney, NSW, Australia: Association for Computing Machinery, pp. 1849–1858. ISBN: 978-1-4503-3664-2. DOI: 10.1145/2783258.2788583.

Holtz, Lobel, Liskovich, and Aral (2020). "Reducing interference bias in online marketplace pricing experiments". *arXiv preprint arXiv:2004.12489*.

Hopkins (2020). *Increasing experimental power with variance reduction at the BBC — by Frank Hopkins — BBC Data Science — Medium*. https://medium.com/bbc-data-science/increasir (Accessed on 02/25/2022).

Howard and Ramdas (2019). "Sequential estimation of quantiles with applications to A/B-testing and best-arm identification". *arXiv preprint arXiv:1906.09712*.

Imai and Ratkovic (2013). "Estimating treatment effect heterogeneity in randomized program evaluation". *The Annals of Applied Statistics* 7.1, pp. 443–470. DOI: 10.1214/12-AOAS593.

Imbens, Kallus, Mao, and Wang (2022). *Long-term Causal Inference Under Persistent Confounding via Data Combination*. DOI: 10.48550/ARXIV.2202.07234.

Isaac (2021). *Facebook Wrestles With the Features It Used to Define Social Networking*.

Issa Mattos, Bosch, and Olsson (2019). "Multi-armed bandits in the wild: Pitfalls and strategies in online experiments". *Information and Software Technology* 113, pp. 68–81. ISSN: 0950-5849. DOI: 10.1016/j.infsof.2019.05.004.

Ivaniuk (2020). *Our evolution towards T-REX: The prehistory of experimentation infrastructure at LinkedIn — LinkedIn Engineering*. https://engineering.linkedin.com/blog/2020/our-evol (Accessed on 02/14/2022).

Jackson (2018). *How Booking.com increases the power of online experiments with CUPED — Booking.com Data Science*. https://booking.ai/how-booking-com-increases-the-power-of-onli (Accessed on 01/13/2021).

Johari, Koomen, Pekelis, and Walsh (2017). "Peeking at A/B Tests: Why it matters, and what to do about it". *Proceedings of the 23rd ACM SIGKDD International Conference on*

*Knowledge Discovery and Data Mining*. KDD '17. Halifax, NS, Canada: Association for Computing Machinery, pp. 1517–1525. ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3097992.

Johari, Pekelis, and Walsh (2019). "Always Valid Inference: Bringing Sequential Analysis to A/B Testing". *arXiv:1512.04922 [math, stat]*. arXiv: 1512.04922.

Ju, Hu, Henderson, and Hong (2019). "A Sequential Test for Selecting the Better Variant: Online A/B Testing, Adaptive Allocation, and Continuous Monitoring". *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. WSDM '19. Melbourne VIC, Australia: Association for Computing Machinery, pp. 492–500. ISBN: 9781450359405. DOI: 10.1145/3289600.3291025.

Kamalbasha and Eugster (2021). "Bayesian A/B testing for business decisions". *Data science–analytics and applications*. Springer, pp. 50–57.

Karrer, Shi, Bhole, Goldman, Palmer, Gelman, Konutgan, and Sun (2021). "Network experimentation at scale". *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3106–3116.

Katzir, Liberty, and Somekh (2012). "Framework and algorithms for network bucket testing". *Proceedings of the 21st international conference on World Wide Web*. WWW '12. New York, NY, USA: Association for Computing Machinery, pp. 1029–1036. ISBN: 978-1-4503-1229-5. DOI: 10.1145/2187836.2187974.

Keenan (2022). *Global Ecommerce Explained: Stats and Trends to Watch in 2022*. https://www.shopify.ca (Accessed on 09/16/2022).

Kemp (2022). *DIGITAL 2022: Global Overview Report*. https://datareportal.com/reports/digital-20 (Accessed on 09/16/2022).

Kharitonov, Drutsa, and Serdyukov (2017). "Learning Sensitive Combinations of A/B Test Metrics". *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. WSDM '17. Cambridge, United Kingdom: Association for Computing Machinery, pp. 651–659. ISBN: 9781450346757. DOI: 10.1145/3018661.3018708.

Kharitonov, Vorobev, Macdonald, Serdyukov, and Ounis (2015). "Sequential Testing for Early Stopping of Online Experiments". *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '15. Santiago, Chile: Association for Computing Machinery, pp. 473–482. ISBN: 9781450336215. DOI: 10.1145/2766462.2767729.

Kohavi (2012). "Online controlled experiments: introduction, learnings, and humbling statistics". *Proceedings of the sixth ACM conference on Recommender systems*. RecSys '12. New York, NY, USA: Association for Computing Machinery, pp. 1–2. ISBN: 978-1-4503-1270-7. DOI: 10.1145/2365952.2365954.

Kohavi, Deng, and Vermeer (2022). "A/B Testing Intuition Busters: Common Misunderstandings in Online Controlled Experiments". DOI: 10.1145/3534678.3539160.

Kohavi (2022). *Build vs Buy*. https://bit.ly/RKABClassBuildVsBuyDeck.

Kohavi, Deng, Frasca, Longbotham, Walker, and Xu (2012). "Trustworthy online controlled experiments: five puzzling outcomes explained". *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '12. Beijing, China: Association for Computing Machinery, pp. 786–794. ISBN: 978-1-4503-1462-6. DOI: 10.1145/2339530.2339653.

Kohavi, Deng, Frasca, Walker, Xu, and Pohlmann (2013). "Online controlled experiments at large scale". *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*. the 19th ACM SIGKDD international conference. Chicago, Illinois, USA: ACM Press, p. 1168. ISBN: 978-1-4503-2174-7. DOI: 10.1145/2487575.2488

Kohavi, Deng, Longbotham, and Xu (2014). "Seven rules of thumb for web site experimenters". *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*. the 20th ACM SIGKDD international conference. New York, New York, USA: ACM Press, pp. 1857–1866. ISBN: 978-1-4503-2956-9. DOI: 10.1145/2623330.2623341.

Kohavi, Longbotham, Sommerfield, and Henne (2009). "Controlled experiments on the web: survey and practical guide". *Data Mining and Knowledge Discovery* 18.1, pp. 140–181. ISSN: 1384-5810, 1573-756X. DOI: 10.1007/s10618-008-0114-1.

Kohavi, Tang, and Xu (2020). *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge: Cambridge University Press. ISBN: 978-1-108-72426-5. DOI: 10.1017/9781108653985.

Kohavi and Thomke (2017). "The Surprising Power of Online Experiments". *Harvard Business Review* (September–October 2017). ISSN: 0017-8012.

Kohlmeier (2022). *Microsoft's Experimentation Platform: How We Build a World Class Product - Microsoft Research*. https://www.microsoft.com/en-us/research/group/experimentatio (Accessed on 02/14/2022).

Koning, Hasan, and Chatterji (2022). "Experimentation and Start-up Performance: Evidence from A/B Testing". *Management Science*.

Kontotasiou (2021). *The Guide to Ethical A/B Testing: The Missing Component of Your Optimization Program*. convert.com/blog/a-b-testing/ethical-ab-testing-guide/.

Koutra (2017). "Designing experiments on networks". PhD thesis. University of Southampton. 222 pp.

Kramer, Guillory, and Hancock (2014). "Experimental evidence of massive-scale emotional contagion through social networks". *Proceedings of the National Academy of Sciences* 111.24, pp. 8788–8790.

Lan, Bakthavachalam, Sharan, Douriez, Azarnoush, and Kroll (2022). *A Survey of Causal Inference Applications at Netflix — by Netflix Technology Blog*. https://netflixtechblog.com/a-survey (Accessed on 08/18/2022).

Leskovec, Lang, and Mahoney (2010). "Empirical comparison of algorithms for network community detection". *Proceedings of the 19th international conference on World wide web*. WWW '10. New York, NY, USA: Association for Computing Machinery, pp. 631–640. ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772755.

Letham, Karrer, Ottoni, and Bakshy (2019). "Constrained Bayesian Optimization with Noisy Experiments". *Bayesian Anal.* 14.2, pp. 495–519. DOI: 10.1214/18-BA1110.

Li, Chu, Langford, and Schapire (2010). "A contextual-bandit approach to personalized news article recommendation". *Proceedings of the 19th international conference on World wide web*. WWW '10. New York, NY, USA: Association for Computing Machinery, pp. 661–670. ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772758.

Lindon and Malek (2020). *Anytime-Valid Inference for Multinomial Count Data*. DOI: 10.48550/ARXIV.201

Lindon, Sanden, and Shirikian (2022). "Rapid Regression Detection in Software Deployments through Sequential Testing". *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '22. Washington DC, USA: Association for Computing Machinery, pp. 3336–3346. ISBN: 9781450393850. DOI: 10.1145/3534678.3539099.

Liou and Taylor (2020). "Variance-Weighted Estimators to Improve Sensitivity in Online Experiments". *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 837–850.

Liu, Mao, and Kang (2021). "Trustworthy and Powerful Online Marketplace Experimentation with Budget-split Design". *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3319–3329.

Liu, Sun, Varshney, and Xu (2019). "Large-scale online experimentation with quantile metrics". *arXiv preprint arXiv:1903.08762*.

Liu, Mandel, Brunskill, and Popovic (2014). "Trading Off Scientific Knowledge and User Learning with Multi-Armed Bandits". *EDM*.

Lomas, Forlizzi, Poonwala, Patel, Shodhan, Patel, Koedinger, and Brunskill (2016). "Interface Design Optimization as a Multi-Armed Bandit Problem". *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. New York, NY, USA: Association for Computing Machinery, pp. 4142–4153. ISBN: 978-1-4503-3362-7. DOI: 10.1145/2858036.2858425.

Luca and Bazerman (2021). *The power of experiments: Decision making in a data-driven world*. Mit Press.

Lux (2018). *Analyzing Experiment Outcomes: Beyond Average Treatment Effects - Uber Engineering Blog*. https://eng.uber.com/analyzing-experiment-outcomes/. (Accessed on 03/21/2022).

Manzi (2012). *UNCONTROLLED The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*.

McFarland (2012). *Experiment!: Website conversion rate optimization with A/B and multivariate testing*. New Riders. 190 pp. ISBN: 978-0-13-304008-1.

Mucha, Richardson, Macon, Porter, and Onnela (2010). "Community Structure in Time-Dependent, Multiscale, and Multiplex Networks". *Science* 328.5980, pp. 876–878. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1184819. arXiv: 0911.1824.

Nandy, Basu, Chatterjee, and Tu (2020). "A/B testing in dense large-scale networks: design and inference". *Advances in Neural Information Processing Systems* 33, pp. 2870–2880.

Newman (2006). "Modularity and community structure in networks". *Proceedings of the National Academy of Sciences* 103.23. Publisher: National Academy of Sciences Section: Physical Sciences, pp. 8577–8582. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0601602103.

Parker, Gilmour, and Schormans (2017). "Optimal design of experiments on connected units with application to social networks". *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 66.3. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/rssc.12170, pp. 455–480. ISSN: 1467-9876. DOI: 10.1111/rssc.12170.

Parks, Aurisset, and Ramm (2017). *Innovating Faster on Personalization Algorithms at Netflix Using Interleaving*. https://netflixtechblog.com/interleaving-in-online-experiments-at-

Pekelis (2015). *Statistics for the Internet Age: The Story Behind Optimizely's New Stats Engine*. https://www.optimizely.com/insights/blog/statistics-for-the-internet-age-the-sto
(Accessed on 03/08/2022).

Petersen, Witten, and Simon (2016). "Fused lasso additive model". *Journal of Computational and Graphical Statistics* 25.4, pp. 1005–1025.

Peysakhovich and Lada (2016). "Combining observational and experimental data to find heterogeneous treatment effects". *arXiv preprint arXiv:1611.02385*.

Pokhiko, Zhang, Kang, and Mays (2019). "D-optimal Design for Network A/B Testing". *Journal of Statistical Theory and Practice* 13.4, p. 61. ISSN: 1559-8608, 1559-8616. DOI: 10.1007/s42519-019-0058-3. arXiv: 1902.00482.

Poyarkov, Drutsa, Khalyavin, Gusev, and Serdyukov (2016). "Boosted Decision Tree Regression Adjustment for Variance Reduction in Online Controlled Experiments". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, pp. 235–244. ISBN: 9781450342322. DOI: 10.1145/2939672.2939688.

Pramanik, Johnson, and Bhattacharya (2021). "A modified sequential probability ratio test". *Journal of Mathematical Psychology* 101, p. 102505.

Prentice (1989). "Surrogate endpoints in clinical trials: definition and operational criteria". *Statistics in medicine* 8.4, pp. 431–440.

Radlinski and Craswell (2013). "Optimized Interleaving for Online Retrieval Evaluation". *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. WSDM '13. Rome, Italy: Association for Computing Machinery, pp. 245–254. ISBN: 9781450318693. DOI: 10.1145/2433396.2433429.

Rajkumar, Saint-Jacques, Bojinov, Brynjolfsson, and Aral (2022). "A causal test of the strength of weak ties". *Science* 377.6612, pp. 1304–1310.

Robinson (1988). "Root-N-Consistent Semiparametric Regression". *Econometrica* 56.4, pp. 931–954. ISSN: 00129682, 14680262.

Rosenbaum and Rubin (1983). "The central role of the propensity score in observational studies for causal effects". *Biometrika* 70.1, pp. 41–55.

Rudder (2014). *We experiment on human beings!* https://www.gwern.net/docs/psychology/okcupid/we

Sadeghi, Chien, and Arora (2019). "Sliced Designs for Multi-Platform Online Experiments". *Technometrics*. Publisher: Taylor & Francis, pp. 1–16. ISSN: 0040-1706. DOI: 10.1080/00401706.2019.1

Sadeghi, Gupta, Gramatovici, Lu, Ai, and Zhang (2022). "Novelty and primacy: a long-term estimator for online experiments". *Technometrics* 64.4, pp. 524–534.

Saint-Jacques (2019). *Detecting interference: An A/B test of A/B tests — LinkedIn Engineering*. https://engineering.linkedin.com/blog/2019/06/detecting-interference--an-a-b-t (Accessed on 02/22/2022).

Saint-Jacques, Varshney, Simpson, and Xu (2019). "Using Ego-Clusters to Measure Network Effects at LinkedIn". *arXiv preprint arXiv:1903.08755*.

Sangho Yoon (2018). *Designing A/B tests in a collaboration network*. The Unofficial Google Data Science Blog. Library Catalog: www.unofficialgoogledatascience.com. URL: http://www.unofficia (visited on 06/11/2020).

Saveski, Pouget-Abadie, Saint-Jacques, Duan, Ghosh, Xu, and Airoldi (2017). "Detecting Network Effects: Randomizing Over Randomized Experiments". *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '17. Halifax, NS, Canada: Association for Computing Machinery, pp. 1027–1035. ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3098192.

Schroeder (2021). *The Data Analytics Profession And Employment Is Exploding: Three Trends That Matter*. https://www.forbes.com/sites/bernhardschroeder/2021/06/11/the-data-a (Accessed on 03/10/2022).

Scott (2010). "A modern Bayesian look at the multi-armed bandit". *Applied Stochastic Models in Business and Industry* 26.6, pp. 639–658. ISSN: 1524-1904. DOI: 10.1002/asmb.874.

— (2015). "Multi-armed bandit experiments in the online service economy". *Applied Stochastic Models in Business and Industry* 31.1, pp. 37–45. ISSN: 1524-1904. DOI: 10.1002/asmb.2104.

Sexauer (2022). *CUPED on Statsig*. https://blog.statsig.com/cuped-on-statsig-d57f23122d0e. (Accessed on 08/18/2022).

Sharma (2021). *Reducing Experiment Durations - Eppo Blog*. https://www.geteppo.com/blog/reducing- (Accessed on 02/25/2022).

— (2022). *Bending time in experimentation - Eppo Blog*. https://www.geteppo.com/blog/bending-time (Accessed on 08/18/2022).

Shi, Wang, Luo, Song, Zhu, and Ye (2020). "A Reinforcement Learning Framework for Time-Dependent Causal Effects Evaluation in A/B Testing". *arXiv preprint arXiv:2002.01711*.

Spang, Hannan, Kunamalla, Huang, McKeown, and Johari (2021). "Unbiased experiments in congested networks". *Proceedings of the 21st ACM Internet Measurement Conference*, pp. 80–95.

Stanley, Shai, Taylor, and Mucha (2016). "Clustering Network Layers with the Strata Multilayer Stochastic Block Model". *IEEE Transactions on Network Science and Engineering* 3.2. Conference Name: IEEE Transactions on Network Science and Engineering, pp. 95–105. ISSN: 2327-4697. DOI: 10.1109/TNSE.2016.2537545.

Stevens and Hagar (2022). "Comparative Probability Metrics: Using Posterior Probabilities to Account for Practical Equivalence in A/B tests". *The American Statistician* 76.3, pp. 224–237.

Stucchio (2015). "Bayesian A/B Testing at VWO", p. 33.

Syrgkanis, Lei, Oprescu, Hei, Battocchi, and Lewis (2019). "Machine learning estimation of heterogeneous treatment effects with instruments". *Advances in Neural Information Processing Systems*, pp. 15193–15202.

Tang, Agarwal, O'Brien, and Meyer (2010). "Overlapping experiment infrastructure: more, better, faster experimentation". *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '10. Washington, DC, USA: Association for Computing Machinery, pp. 17–26. ISBN: 978-1-4503-0055-1. DOI: 10.1145/1835804.1835810.

Thomke (2020). *Experimentation works: The surprising power of business experiments*. Harvard Business Press.

Tran and Zheleva (2019). "Learning triggers for heterogeneous treatment effects". *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 5183–5190.

Tsiatis (2006). "Semiparametric theory and missing data".

Ugander, Karrer, Backstrom, and Kleinberg (2013). "Graph cluster randomization: network exposure to multiple universes". *arXiv:1305.6979 [physics, stat]*. arXiv: 1305.6979.

Urban, Sreenivasan, and Kannan (2016). *It's All A/Bout Testing: The Netflix Experimentation Platform — by Netflix Technology Blog — Netflix TechBlog*. https://netflixtechblog.com/its-a (Accessed on 10/26/2021).

Visser (2020). *In-house experimentation platforms*. https://www.linkedin.com/pulse/in-house-experi (Accessed on 09/15/2022).

Von Ahn (2022). *Shareholder Letter Q2 2022*. https://investors.duolingo.com/static-files/ae55dd3

Wager and Athey (2018). "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". *Journal of the American Statistical Association* 113.523, pp. 1228–1242. DOI: 10.1080/01621459.2017.1319839. eprint: https://doi.org/10.1080/01621459.2017.131

Wald (1945). "Sequential Tests of Statistical Hypotheses". *The Annals of Mathematical Statistics* 16.2, pp. 117–186. ISSN: 00034851.

Wald (1947). *Sequential analysis*. Courier Corporation.

Walker and Muchnik (2014). "Design of Randomized Experiments in Networks". *Proceedings of the IEEE* 102.12. Conference Name: Proceedings of the IEEE, pp. 1940–1951. ISSN: 1558-2256. DOI: 10.1109/JPROC.2014.2363674.

Wang, Gupta, Lu, Mahmoudzadeh, and Liu (2019). "On Heavy-user Bias in A/B Testing". *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 2425–2428.

Wildman (2019). *Using A/B Testing, Factorial Design, and Multivariate Tests for Deep Visitor Insights*. Library Catalog: www.thecreativemomentum.com. URL: https://www.thecreativemoment (visited on 07/13/2020).

Xie and Aurisset (2016). "Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, pp. 645–654. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939733.

Xie, Chen, and Shi (2018). "False Discovery Rate Controlled Heterogeneous Treatment Effect Detection for Online Controlled Experiments". *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '18. London, United Kingdom: Association for Computing Machinery, pp. 876–885. ISBN: 978-1-4503-5552-0. DOI: 10.1145/3219819.3219860.

Xu, Chen, Fernandez, Sinno, and Bhasin (2015). "From Infrastructure to Culture: A/B Testing Challenges in Large Scale Social Networks". *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*. the 21th ACM SIGKDD International Conference. Sydney, NSW, Australia: ACM Press, pp. 2227–2236. ISBN: 978-1-4503-3664-2. DOI: 10.1145/2783258.2788602.

Xu, Duan, and Huang (2018). "SQR: Balancing Speed, Quality and Risk in Online Experiments". *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. London United Kingdom: ACM, pp. 895–904. ISBN: 978-1-4503-5552-0. DOI: 10.1145/3219819.3219875.

Yu, Lu, and Song (2020). *A New Framework for Online Testing of Heterogeneous Treatment Effect*. arXiv: 2002.03277 [stat.ME].

Zhang, Coey, Goldman, and Karrer (2021). "Regression Adjustment with Synthetic Controls in Online Experiments".

Zhang, Du, Andersen, and Hel (2022). *Beyond A/B Test: Speeding up Airbnb Search Ranking Experimentation through Interleaving*. https://medium.com/airbnb-engineering/beyond-a-b-test-

Zhang and Kang (2020). "Optimal Design for A/B Testing in the Presence of Covariates and Network Connection". *arXiv:2008.06476 [stat]*. arXiv: 2008.06476.

Zhao, Zeng, Rush, and Kosorok (2012). "Estimating Individualized Treatment Rules Using Outcome Weighted Learning". *Journal of the American Statistical Association* 107.499. PMID: 23630406, pp. 1106–1118. DOI: 10.1080/01621459.2012.695674. eprint: https://doi.org/10.

Zhou, Liu, Li, and Hu (2020). "Cluster-Adaptive Network A/B Testing: From Randomization to Estimation". *arXiv:2008.08648 [stat]*. arXiv: 2008.08648.