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

MANAGEMENT SCIENCE
Vol. 69, No. 7, July 2023, pp. 3759–3777
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Design and Analysis of Switchback Experiments

Iavor Bojinov,[a] David Simchi-Levi,[b] Jinglong Zhao[c,*]

[a] Technology and Operations Management Unit, Harvard Business School, Boston, Massachusetts 02163; [b] Institute for Data, Systems, and Society, Department of Civil and Environmental Engineering, and Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139; [c] Operations and Technology Management Department, Questrom School of Business, Boston University, Boston, Massachusetts 02215

*Corresponding author

Contact: ibojinov@hbs.edu, https://orcid.org/0000-0002-3470-8539 (IB); dslevi@mit.edu, https://orcid.org/0000-0002-4650-1519 (DS-L); jinglong@bu.edu, https://orcid.org/0000-0003-0986-0085 (JZ)

Received: September 25, 2020
Revised: May 24, 2021; January 4, 2022
Accepted: March 1, 2022
Published Online in Articles in Advance: November 1, 2022

https://doi.org/10.1287/mnsc.2022.4583

Abstract. Switchback experiments, where a firm sequentially exposes an experimental unit to random treatments, are among the most prevalent designs used in the technology sector, with applications ranging from ride-hailing platforms to online marketplaces. Although practitioners have widely adopted this technique, the derivation of the optimal design has been elusive, hindering practitioners from drawing valid causal conclusions with enough statistical power. We address this limitation by deriving the optimal design of switchback experiments under a range of different assumptions on the order of the carryover effect—the length of time a treatment persists in impacting the outcome. We cast the optimal experimental design problem as a minimax discrete optimization problem, identify the worst-case adversarial strategy, establish structural results, and solve the reduced problem via a continuous relaxation. For switchback experiments conducted under the optimal design, we provide two approaches for performing inference. The first provides exact randomization-based p-values, and the second uses a new finite population central limit theorem to conduct conservative hypothesis tests and build confidence intervals. We further provide theoretical results when the order of the carryover effect is misspecified and provide a data-driven procedure to identify the order of the carryover effect. We conduct extensive simulations to study the numerical performance and empirical properties of our results and conclude with practical suggestions.

History: Accepted by George Shanthikumar, big data analytics.
Funding: The authors thank the Massachusetts Institute of Technology (MIT)-IBM partnership in Artificial Intelligence and the MIT Data Science Laboratory for support.
Supplemental Material: Data and the online appendix are available at https://doi.org/10.1287/mnsc.2022.4583.

Keywords: design of experiments • discrete optimization • central limit theorem • switchback experiments

1. Introduction

Academic scholars have appreciated the benefits that experimentation brings to firms for many decades (March 1991, Sitkin 1992, Sarasvathy 2001, Thomke 2001, Johari et al. 2015, Kohavi and Thomke 2017, Sun et al. 2018, Xiong et al. 2019). However, widespread adoption of the practice has only taken off in the last decade, partly fueled by the rapid cost reductions achieved by firms in the technology sector (Kohavi et al. 2007, 2009; Bakshy et al. 2014; Azevedo et al. 2019; Kohavi et al. 2020). Most large firms now possess internal tools for experimentation, and a growing number of smaller and more conventional companies are purchasing the capabilities from third-party sellers that offer full-stack integration (Thomke 2020). These tools typically allow simple "A/B" tests that compare the standard offering "A" to a new or improved version "B." The comparisons are made across a range of different business outcomes, and the tests are usually conducted for at least a week (Kohavi et al. 2020). This simple practice has provided tremendous value to firms (Koning et al. 2019).

However, some firms and authors have recognized the limitations of these simple A/B tests (Gupta et al. 2019, Bojinov et al. 2020); the two most prominent being handling interference (the scenario where the assignment of one subject impacts another's outcomes) and estimating heterogeneous (or personalized) effects. For example, many online platforms and retail marketplaces often observe varying levels of interference when conducting experiments (see Chamandy 2016, Cui et al. 2020, Kastelman and Ramesh 2018, Farronato et al. 2018, Glynn et al. 2020, Holtz et al. 2020, Li et al. 2021 for online platforms like Airbnb, DoorDash, Lyft, and Uber and Caro and Gallien 2012, Ferreira et al. 2016, Cui et al.

2019, and Ma et al. 2021 for retail markets like Amazon, AB InBev, Rue la la, Zara) and desire to estimate heterogeneous effects (see Nie et al. 2018, Deshpande et al. 2018, McFowland et al. 2018, Hadad et al. 2019).

In this paper, we simultaneously tackle both of these two challenges by developing a theoretical framework for the optimal design and analysis of switchback experiments under the minimal amount of assumptions. In switchback experiments, we sequentially expose a unit to a random treatment, measure its response, and repeat the procedure for a fixed period of time (Robins 1986, Bojinov and Shephard 2019). By administering alternate treatments to the same unit, we can directly estimate an individual level causal effect and alleviate the challenges posed by interference.

In addressing the two challenges, many works in the literature assume specific outcome models under interference. Wager and Xu (2019), Johari et al. (2020), and Li et al. (2021) work on experimental design for two-sided online platforms by assuming that the interference can be captured via game-theoretic modeling. Glynn et al. (2020) assumes an underlying Markov chain model and formulates the experimental design problem as estimating the difference between two steady state reward distributions. Some other literature directly models the interference through a network, for example, Li et al. (2015), Eckles et al. (2017), Sussman and Airoldi (2017), Athey et al. (2018), Basse et al. (2019a), Puelz et al. (2019). In such models, a treatment assigned to one node of the network creates a "spillover effect," which impacts the outcomes of the neighboring nodes. All of the above methods make specific assumptions on the outcome models. If these assumptions hold, the above methods correctly identify the causal effects (or the model parameters) with great precision; if these assumptions do not hold, the estimates are likely biased.

Unlike the above works, we make no specific outcome model assumptions in this paper. Instead, we make assumptions about the existence of the carryover effects, which refer to the persistence of past interventions in impacting the future outcomes. More specifically, we make assumptions on the order of carryover effects, which refers to the duration of time periods of such persistence. We then establish formal results on the optimal design of switchback experiments under different assumptions of the order of the carryover effects; we also propose a data-driven procedure to estimate the order of the carryover effects.

1.1. Applications

There are two classes of applications where switchback experiments are widely used in practice. The first arises when units interfere with each other either through a network or some more complicated unknown structure. For example, consider a ride-hailing platform that wants to test a new fare pricing algorithm's effectiveness in a large city (Farronato et al. 2018). Administering the test version to a subset of drivers can impact their behavior, which, in turn, could change the behavior of drivers that are receiving the old version. Directly comparing the revenue generated by the drivers across the two groups will likely provide a biased estimate of what would happen if everyone were assigned to the new version compared with the old. Instead, practitioners treat the city as a single aggregated unit and use a switchback experiment to estimate the intervention's effectiveness, thereby alleviating the problem caused by interference. A similar issue often arises in revenue management when, for example, a retailer wants to test the effectiveness of a new promotion planning algorithm (Ferreira et al. 2016). Administering the new version to a subset of stock keeping units (SKUs) cannibalizes the sales from the other SKUs. Again comparing the generated revenue across the two groups is unlikely to provide an accurate measure of the promotion's effectiveness. Again, practitioners treat all the SKUs as a single aggregated unit and use a switchback experiment to obtain accurate estimates of the promotion's effectiveness.

The second application arises when we have a limited number of experimental units, and we believe the effects are likely to be heterogeneous. For example, Bojinov and Shephard (2019) used switchback experiments to make causal claims about the relative effectiveness of algorithms compared with humans at executing large financial trades across a range of financial markets. More generally, psychologists and biostatisticians regularly use switchback experiments whenever studying the effectiveness of an intervention on a single unit, for example, Lillie et al. (2011) and Boruvka et al. (2018).

1.2. Main Contributions

There are three significant challenges to using switchback experiments. The first is that causal estimators from switchback experiments have large variances as the precision is a function of the total number of assignments. The second is that past interventions are likely to impact future outcomes; this is often referred to as a carryover effect. Typically, many authors assume that there are no carryover effects (Chamberlain 1982, Athey and Imbens 2018, Imai and Kim 2019), although some recent work has relaxed this assumption (Robins 1986, Sobel 2012, Bojinov et al. 2021). The third is that standard super population inference—where researchers either assume a model for the outcome or that the units are sampled from an infinitely large population—requires unrealistic assumptions that fail to capture the problem's personalized nature (Bojinov and Shephard 2019).

This paper's main contributions are to address these three challenges and present a framework that allows

firms and researchers to run reliable switchback experiments. First, we derive optimal designs for switchback experiments, ensuring that we select a design that leads to the lowest variance among the most popular class assignment mechanisms. The designs are optimal in the sense that we search for both the optimal randomization points and the optimal randomization probabilities, which, together, capture the most general class of randomization mechanisms. Second, we assume the presence of a carryover effect and show that our estimation and inference are valid both when the order of the carryover effect is correctly specified and misspecified, the latter leading to a minor increase in the variance. For practitioners, we also propose a method to identify the order of the carryover effect by running a series of carefully designed switchback experiments. Finally, we take a purely design-based perspective on uncertainty; that is, we treat the outcomes as unknown but fixed (or equivalently, we condition on the set of potential outcomes) and assume that the assignment mechanism is the only source of randomness (Fisher et al. 1937, Kempthorne 1955, Rubin 1980, Abadie et al. 2020). The main benefit of a design-based perspective is that the inference, and in turn the causal conclusions, do not depend on our ability to correctly specify a model describing the phenomena we are studying, ensuring that our findings are wholly nonparametric and robust to model misspecification (Imbens and Rubin 2015, chapter 5).

1.3. Roadmap

The paper is structured as follows. In Section 2, we define the notations, the assumptions, and the assignment mechanism that we focus on, which we will refer to as the regular switchback experiments. In Section 3, we discuss how to design an effective regular switchback experiment under the minimax rule. The design is optimal with respect to (i) the optimal treatment assignment probability and (ii) the randomization frequency and randomization points. We cast the design problem as a minimax discrete optimization problem, identify the worst-case adversarial strategy, establish structural results, and then explicitly find the optimal design. In Section 4, we discuss how to perform inference and conduct statistical testing based on the results obtained from an optimally designed switchback experiment. We propose an exact test for sharp null hypotheses and an asymptotic test for testing the average treatment effect. We also discuss how to make an inference when the carryover effect is misspecified and how to conduct hypothesis testing to identify the true order of the carryover effect. In Section 5, we run simulations to test the correctness and effectiveness of our proposed theoretical results under various simulation setups. In Section 6, we give empirical illustrations on how to conduct a switchback experiment in practice and conclude with limitations that may lead to future research directions. All technical proofs are in the online appendix.

2. Notations, Assumptions, and Regular Switchback Experiments

2.1. Assignment Paths and Potential Outcomes

We focus our discussion on a single experimental unit. For example, this unit could be a ride-hailing platform testing the effectiveness of a new fare pricing algorithm in a city. At each time point $t \in [T] = \{1, 2, \ldots, T\}$, we assign the unit to receive an intervention $W_t \in \{0, 1\}$. For example, one experimental period could be one to two hours for a ride-hailing platform and T could be two weeks, that is, $T = 336$ when one period is one hour. In some applications, the time horizon T is predetermined, for example, a typical experimental duration for a ride-hailing platform is a few weeks; however, when T is not predetermined, Section 6 provides details for how to choose an appropriate T. Throughout most of this paper, with the exception being the derivation of our asymptotic results, we consider T to be a known, fixed constant.

Following convention, we say that the unit is assigned to treatment if $W_t = 1$ and control when $W_t = 0$; in A/B testing terminology, "A" is control and "B" is treatment. For example, Chamandy (2016) studied how a new surge-pricing subsidy (the treatment) compared with the current setup without the subsidy (the control). The assignment path is then the collection of assignments and is denoted using a vector notation whose dimensions are specified in the subscript, $\boldsymbol{W}_{1:T} = (W_1, W_2, \ldots, W_T) \in \{0, 1\}^T$. We adopt the convention that $\boldsymbol{W}_{1:T}$ stands for a random assignment path, whereas $\boldsymbol{w}_{1:T}$ stands for one realization.

After administering the assigned intervention, we observe a corresponding outcome. For example, this could be the average ride-matching rate (often defined as the proportion of requested rides that were successfully matched with a driver) during each two-hour experimental period. Following the extended potential outcomes framework, at time $t \in [T]$, we posit that for each possible assignment path $\boldsymbol{w}_{1:T}$, there exists a corresponding potential outcome denoted by $Y_t(\boldsymbol{w}_{1:T})$; the set of all potential outcomes are collected in

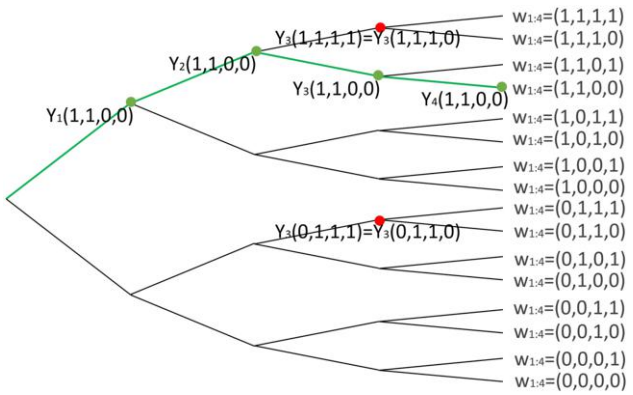
$$\mathbb{Y} = \{Y_t(\boldsymbol{w}_{1:T})\}_{t \in [T], \boldsymbol{w}_{1:T} \in \{0,1\}^T}$$

with support $\mathbb{Y} \in \mathcal{Y}$.

Example 1. When $T = 4$, there are 16 assignment paths as shown in Figure 1. Associated with each assignment path $\boldsymbol{w}_{1:4}$ are four potential outcomes $Y_1(\boldsymbol{w}_{1:4}), Y_2(\boldsymbol{w}_{1:4}), Y_3(\boldsymbol{w}_{1:4}), Y_4(\boldsymbol{w}_{1:4})$.

Throughout this paper, we do not directly model the potential outcomes or impose a parametric relationship with the assignment path; instead, we treat them as

Figure 1. (Color online) Illustrator of Assignment Paths and Potential Outcomes When $T = 4$



Notes. The green path stands for one assignment path $\boldsymbol{w}_{1:4} = (1,1,0,0)$. Following the green path, there are four potential outcomes. The two red dots each stand for two potential outcomes that are equal under Assumption 1. And the potential outcomes at the two red dots are equal if Assumption 2 is further assumed.

unknown but fixed quantities or, equivalently, we implicitly condition on $\mathbb{Y}$. Our setup does not preclude the possibility that the potential outcomes were generated through a dynamic process; however, it allows us to be completely agnostic to the data-generating process, making our causal claims more objective. To make inference possible, we rely on the variation introduced by the random assignment path; this is commonly referred to as finite-sample or design-based perspective and has a long history going back to Fisher et al. (1937), Kempthorne (1955), Rubin (1980), and Neyman et al. (1990). Unlike traditional sampling-based inference, the design-based approach does not require a hypothetical population from which to sample experimental units; see Imbens and Rubin (2015) and Abadie et al. (2020) for recent reviews. Instead, we make two assumptions that limit the dependence of the potential outcomes on assignment paths. Below let $\{t : t'\} = \{t, t+1, \ldots, t'\}$, for any $t < t' \in [T]$.

Assumption 1 (Nonanticipating Potential Outcomes). *For any* $t \in [T]$, $\boldsymbol{w}_{1:t} \in \{0,1\}^t$, *and for any* $\boldsymbol{w}'_{t+1:T}, \boldsymbol{w}''_{t+1:T} \in \{0,1\}^{T-t}$,

$$Y_t(\boldsymbol{w}_{1:t}, \boldsymbol{w}'_{t+1:T}) = Y_t(\boldsymbol{w}_{1:t}, \boldsymbol{w}''_{t+1:T}).$$

Assumption 1 states that the potential outcomes at time t do not depend on future treatments (Basse et al. 2019b, Bojinov and Shephard 2019, Rambachan and Shephard 2019). Because we control the assignment mechanism instead of letting the experimental units to administer future assignments (e.g., at a ride-hailing platform, a passenger does not know the price in the next hour), the design ensures that this assumption is satisfied.

Example 2 (Example 1 Continued). Under Assumption 1, $Y_3(1,1,1,1) = Y_3(1,1,1,0)$. In Figure 1, the dot at $Y_3(1,1,1)$ stands for both $Y_3(1,1,1,1)$ and $Y_3(1,1,1,0)$.

Assumption 2 (m-Carryover Effects). *There exists a fixed and given* m, *such that for any* $t \in \{m+1, m+2, \ldots, T\}$, $\boldsymbol{w}_{t-m:T} \in \{0,1\}^{T-t+m+1}$, *and for any* $\boldsymbol{w}'_{1:t-m-1}$, $\boldsymbol{w}''_{1:t-m-1} \in \{0,1\}^{t-m-1}$,

$$Y_t(\boldsymbol{w}'_{1:t-m-1}, \boldsymbol{w}_{t-m:T}) = Y_t(\boldsymbol{w}''_{1:t-m-1}, \boldsymbol{w}_{t-m:T}).$$

Assumption 2 restricts the order of the carryover effect (Laird et al. 1992, Senn and Lambrou 1998, Bojinov and Shephard 2019, Basse et al. 2019b). The validity of Assumption 2 depends on the setting and requires practitioners to use their domain knowledge to choose an appropriate m. Examples arise in ride hailing in which the effect of surge pricing on a ride-hailing platform typically dissipates after one or two hours, depending on the city size (Garg and Nazerzadeh 2019). Moreover, in Section 4.4, we propose a data-driven procedure for selecting an appropriate m.

Assumptions 1 and 2 allow us to simplify notation. For any $t \in \{m+1, \ldots, T\}$ and any two assignment paths $\boldsymbol{w}_{1:T}, \boldsymbol{w}'_{1:T} \in \{0,1\}^{m+1}$, whenever $\boldsymbol{w}_{t-m:t} = \boldsymbol{w}'_{t-m:t}$, this leads to

$$Y_t(\boldsymbol{w}_{1:T}) = Y_t(\boldsymbol{w}'_{1:T}).$$

In the remainder of this paper, we will write $Y_t(\boldsymbol{w}_{t-m:t}) := Y_t(\boldsymbol{w}_{1:T})$ to emphasize the dependence on treatments $\boldsymbol{w}_{t-m:t}$. For example, the potential outcomes at the two red dots in Figure 1 are equal, that is, $Y_3(1,1) := Y_3(1,1,1,1) = Y_3(1,1,1,0) = Y_3(0,1,1,1) = Y_3(0,1,1,0)$.

2.2. Causal Effects

In the potential outcomes approach to causal inference, any comparison of potential outcomes has a causal interpretation. In this paper, we focus on a special set of causal estimands that measure the relative effectiveness of persistently assigning a unit to treatment as opposed to control. For any $p \in \{0,1,\ldots,T-1\}$, let $\mathbf{1}_{p+1} = (1,1,\ldots,1)$ be a vector of $(p+1)$ ones; let $\mathbf{0}_{p+1} = (0,0,\ldots,0)$ be a vector of $(p+1)$ zeros. Define the average lag-p causal effect of consecutive treatments on the outcome, for any $p \in \{0,1,\ldots,T-1\}$,

$$\tau_p(\mathbb{Y}) = \frac{1}{T-p}\sum_{t=p+1}^{T}[Y_t(\mathbf{1}_{p+1}) - Y_t(\mathbf{0}_{p+1})]. \tag{1}$$

This estimand captures the effects of permanently deploying a new policy and has been widely studied in the longitudinal experiments since the early work of Robins (1986).

Remark 1. Although we focus on an average causal effect, all of our results and analysis trivially extend to the total causal effect, which does not normalize, that is, $(T-p)\tau_p(\mathbb{Y})$. The optimal design as we will show in Section 3 will remain unchanged.

In our setup, p reflects the experimental designer's knowledge of the order of the carryover effect; see the discussion below Assumption 2. Such a knowledge is

either correct, which we refer to as the perfect knowledge case ($p = m$), or incorrect, which we refer to as the "misspecified" m case[1] ($p \neq m$). In this section, we focus on the $p = m$ case to derive the optimal design; Section 4.3 considers what happens when m is misspecified by discussing the $p \neq m$ case.

The challenge of causal inference on switchback experiments is that we only observe one assignment path. In other words, for each period t, we observe at most either $Y_t(\mathbf{1}_{p+1})$ or $Y_t(\mathbf{0}_{p+1})$ (and sometimes neither). After conducting a switchback experiment, the observed data contain $\boldsymbol{w}_{1:T}^{\text{obs}}$, the realized assignment path, and $Y_t^{\text{obs}} = Y_t(\boldsymbol{w}_{1:T}^{\text{obs}})$, the observed outcome at time t under the realized assignment path $\boldsymbol{w}_{1:T}^{\text{obs}}$. To link the observed and potential outcomes, we assume there is only one version of the treatment[2] and that there is no noncompliance.

2.3. Regular Switchback Experiments

The design of a switchback experiment induces a probabilistic distribution over assignment paths $\boldsymbol{w}_{1:T} \in \{0,1\}^T$. Formally, a design of a switchback experiment is any $\eta : \{0,1\}^T \to [0,1]$ such that

$$\sum_{\boldsymbol{w}_{1:T} \in \{0,1\}^T} \eta(\boldsymbol{w}_{1:T}) = 1, \quad \eta(\boldsymbol{w}_{1:T}) \geq 0, \ \forall \boldsymbol{w}_{1:T} \in \{0,1\}^T.$$

Explicitly, $\eta(\cdot)$ is the underlying discrete distribution of the random assignment path $\boldsymbol{W}_{1:T}$.

In this paper, we narrow our scope to the family of regular switchback experiments. This family of experiments is parameterized by $\mathbb{T}$ and $\mathbb{Q}$, defined as

$$\mathbb{T} = \{t_0 = 1 < t_1 < t_2 < \cdots < t_K\} \subseteq [T],$$

where $K < T$ is a positive integer and $\mathbb{T}$ contains a total of $K + 1$ integers, which is a subset of all the time indices; and

$$\mathbb{Q} = (q_0, q_1, \ldots, q_K) \in (0,1)^{K+1} := \mathcal{Q},$$

where $\mathbb{Q}$ is a vector of $K + 1$ real numbers between (0, 1). For the ease of notations also denote $t_{K+1} = T + 1$, though our time horizon is only T periods.

Definition 1 (Regular Switchback Experiments). For any $\mathbb{T} = \{t_0 = 1 < t_1 < \cdots < t_K\} \subseteq [T]$, and any $\mathbb{Q} = (q_0, q_1, \ldots, q_K) \in (0,1)^{K+1}$, a regular switchback experiment $(\mathbb{T}, \mathbb{Q})$ administers a probabilistic treatment at any time t, given by

$$\Pr(W_t = 1) = q_k, \ \text{if} \ \ t_k \leq t \leq t_{k+1} - 1. \tag{2}$$

In words, the experimental designer jointly decides on a collection of randomization points, which consists of flipping biased coins at each period $t \in \{t_0, \ldots, t_K\}$, as well as a collection of randomization probabilities behind the biased coins, $(q_0, \ldots, q_K)$. If the resulting flip at period t_k is heads, then the experimental designer assigns the unit to treatment during periods $(t_k, t_k + 1, \ldots, t_{k+1} - 1)$ and otherwise, if tails, assigns the unit to control during periods $(t_k, t_k + 1, \ldots, t_{k+1} - 1)$.

Example 3. When $T = 4$, $\mathbb{T} = \{t_0 = 1, t_1 = 3\}$, $\mathbb{Q} = (q_0, q_1) = (1/2, 1/2)$ corresponds to the following design: with probability one-fourth, $\boldsymbol{W}_{1:4} = (1,1,1,1)$; with probability one-fourth, $\boldsymbol{W}_{1:4} = (1,1,0,0)$; with probability one-fourth, $\boldsymbol{W}_{1:4} = (0,0,1,1)$; with probability one-fourth, $\boldsymbol{W}_{1:4} = (0,0,0,0)$. See Figure 2 (left figure) for the four assignment paths that are in the support of the discrete probability distribution.

Example 4. Not all switchback experiments are regular. For example, when $T = 4$, with probability one-fourth, $\boldsymbol{W}_{1:4} = (1,1,1,0)$; with probability one-fourth, $\boldsymbol{W}_{1:4} = (1,0,0,0)$; with probability one-fourth, $\boldsymbol{W}_{1:4} = (0,1,1,1)$; with probability one-fourth, $\boldsymbol{W}_{1:4} = (0,0,0,1)$. See Figure 2 (right figure) for the four assignment paths that are in the support of the discrete probability distribution.

In Section 3, we show that fair coin flipping (i.e., $q_k = 1/2$, $\forall k \in \{0,1,\ldots,K\}$) is indeed optimal, under a mild assumption.[3] The reason behind fair coin flips reflects our limited assumption on the outcome model and the inherent symmetry in the potential outcomes.

Note that we do not consider adaptive treatment assignments as most firms design the entire experiment before the experiment is launched; the treatment assignments are typically not updated based on the observed outcomes. We briefly outline adaptive experimental designs as future extensions in Section 6.

For any regular switchback experiment $(\mathbb{T}, \mathbb{Q})$, we may use $\mathbb{T}$ to refer to the same experiment when $\mathbb{Q}$ is clear from the context. We denote the underlying discrete probability distribution using $\eta_{\mathbb{T},\mathbb{Q}}(\cdot)$. For any $\mathbb{T}$ and $\mathbb{Q}$, the discrete probability distribution has a total of 2^{K+1} many supports. The assignment path is random and follows the discrete probability distribution $\eta_{\mathbb{T},\mathbb{Q}}(\cdot)$:

$$\eta_{\mathbb{T},\mathbb{Q}}(\boldsymbol{w}_{1:T}) = \begin{cases} \prod_{k=0}^{K} \frac{\mathbb{1}\{w_{t_k} = 1\}}{q_{t_k}} \cdot \frac{\mathbb{1}\{w_{t_k} = 0\}}{\bar{q}_{t_k}}, & \text{if} \quad \forall k \in \{0,1,\ldots,K\}, \\ & w_{t_k} = w_{t_k+1} = \cdots = w_{t_{k+1}-1}, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

In the remainder of this paper, unless explicitly noted, all probabilities and expectations are taken with respect to this discrete probability distribution $\eta_{\mathbb{T},\mathbb{Q}}(\cdot)$.

2.4. Estimation

Now that $\eta_{\mathbb{T},\mathbb{Q}}(\cdot)$ is determined, following any realization of the assignment path $\boldsymbol{w}_{1:T}$, we use the Horvitz-Thompson estimator to estimate the causal effect:

$$\widehat{\tau}_p(\eta_{\mathbb{T},\mathbb{Q}}, \boldsymbol{w}_{1:T}, \mathbb{Y}) = \frac{1}{T-p} \sum_{t=p+1}^{T} \left\{ Y_t^{\text{obs}} \frac{\mathbb{1}\{\boldsymbol{w}_{t-p:t} = \mathbf{1}_{p+1}\}}{\Pr(\boldsymbol{W}_{t-p:t} = \mathbf{1}_{p+1})} - Y_t^{\text{obs}} \frac{\mathbb{1}\{\boldsymbol{w}_{t-p:t} = \mathbf{0}_{p+1}\}}{\Pr(\boldsymbol{W}_{t-p:t} = \mathbf{0}_{p+1})} \right\}. \tag{4}$$

We emphasize that the estimator $\widehat{\tau}_p(\cdot,\cdot,\cdot)$ depends on (i) the probability distribution that the assignment path

Figure 2. (Color online) Two Designs

Notes. The blue lines stand for the possible treatment assignments that a design could administer. Left: regular switchback experiment (Example 3); right: irregular switchback experiment (Example 4).

is sampled from, (ii) the realization of the assignment path, and (iii) the set of potential outcomes.

Example 5. Suppose $T = 4, p = m = 1$. Suppose the assignments are probabilistic and $\Pr(W_t = 1) = \Pr(W_t = 0) = 1/2,\ \forall t \in [4]$. With probability one-sixteenth, the green assignment path as in Figure 1 is administered, $\boldsymbol{W}_{1:4} = (1,1,0,0)$. The estimator is then $\widehat{\tau}_1 = \frac{1}{3}\{4Y_2(1,1) + 0 - 4Y_4(0,0)\}$.

Because the assignment path $\boldsymbol{W}_{1:T}$ is random, this Horvitz-Thompson estimator is also random. Moreover, when the assignment path satisfies a regular switchback, the probabilities in the denominator are known. As we will show in Theorem 1, under the optimal design, these probabilities will be multiplicatives of one-half, allowing us to avoid the known stability issues of the Horvitz-Thompson estimator when the probabilities are extreme (either close to zero or close to one). It is well known that the Horvitz-Thompson estimator is unbiased if the treatment and control probabilities are both nonzero.

Proposition 1 (Unbiasedness of the Horvitz-Thompson Estimator). *In a regular switchback experiment, under Assumptions 1 and 2, the Horvitz-Thompson estimator is unbiased for the average lag-p causal effect of consecutive treatments on outcome, that is,*

$$\mathbb{E}[\widehat{\tau}_p(\eta_{\mathbb{T},\mathbb{Q}}, \boldsymbol{W}_{1:T}, \mathbb{Y})] = \tau_p(\mathbb{Y}).$$

The expectation $\mathbb{E}[\cdot]$ is taken with respect to the random assignment $\boldsymbol{W}_{1:T} \sim \eta_{\mathbb{T},\mathbb{Q}}(\cdot)$. When it is obvious, we will compress the subscript in the expectation writing $\mathbb{E}[\cdot]$ to mean $\mathbb{E}_{\boldsymbol{W}_{1:T}\sim\eta_{\mathbb{T},\mathbb{Q}}}[\cdot]$. The proof to Proposition 1 is standard, by checking the expectations. We defer its proof to Section EC.2 in the online appendix.

2.5. Evaluation of Experiments: The Decision-Theoretic Framework

To evaluate the quality of a design of experiment, we adopt the decision-theoretic framework (Berger 2013, Bickel and Doksum 2015). When the random design is $\eta_{\mathbb{T},\mathbb{Q}}(\cdot)$, for any realization of the assignment path $\boldsymbol{w}_{1:T}$ and any set of potential outcomes $\mathbb{Y}$, we define the loss function

$$L(\eta_{\mathbb{T},\mathbb{Q}}, \boldsymbol{w}_{1:T}, \mathbb{Y}) = \left(\widehat{\tau}_p(\eta_{\mathbb{T},\mathbb{Q}}, \boldsymbol{w}_{1:T}, \mathbb{Y}) - \tau_p(\mathbb{Y})\right)^2$$

and the risk function

$$r(\eta_{\mathbb{T},\mathbb{Q}}, \mathbb{Y}) = \sum_{\boldsymbol{w}_{1:T}\in\{0,1\}^T} \eta_{\mathbb{T},\mathbb{Q}}(\boldsymbol{w}_{1:T}) \cdot \left(\widehat{\tau}_p(\eta_{\mathbb{T},\mathbb{Q}}, \boldsymbol{w}_{1:T}, \mathbb{Y}) - \tau_p(\mathbb{Y})\right)^2. \quad (5)$$

Such a risk function quantifies the expected squared difference between our estimand and estimator. Since the estimator is unbiased, the risk function also has a second interpretation: the variance of the estimator. A design with a lower risk is also a design whose estimator has a lower variance.

Example 6 (Examples 3 and 5 Revisited). Suppose $T = 4$ and $p = m = 1$. As in Example 3, $\mathbb{T} = \{1,3\}$. With probability one-fourth, $\boldsymbol{W}_{1:4} = (1,1,0,0)$, $\widehat{\tau}_1(\mathbb{T}) = \frac{1}{3}\{2Y_2(1,1) - 2Y_4(0,0)\}$, $L(\eta_{\mathbb{T},\mathbb{Q}}, \boldsymbol{w}_{1:T}, \mathbb{Y}) = \frac{1}{9}(Y_2(1,1) + Y_2(0,0) - Y_3(1,1) + Y_3(0,0) - Y_4(1,1) - Y_4(0,0))^2$. As in Example 5, $\tilde{\mathbb{T}} = \{1,2,3,4\}$. With probability one-sixteenth, $\boldsymbol{W}_{1:4} = (1,1,0,0)$, $\widehat{\tau}_1(\tilde{\mathbb{T}}) = \frac{1}{3}\{4Y_2(1,1) - 4Y_4(0,0)\}$, $L(\eta_{\tilde{\mathbb{T}},\mathbb{Q}}, \boldsymbol{w}_{1:T}, \mathbb{Y}) = \frac{1}{9}(3Y_2(1,1) + Y_2(0,0) - Y_3(1,1) + Y_3(0,0) - Y_4(1,1) - 3Y_4(0,0))^2$.

Example 6 suggests that, even if the two realizations of the assignment path are the same and the potential outcomes are the same, because the probability distributions $\eta_{\mathbb{T},\mathbb{Q}}$ and $\eta_{\tilde{\mathbb{T}},\mathbb{Q}}$ are distinct, the corresponding estimators $\widehat{\tau}_1(\mathbb{T})$ and $\widehat{\tau}_1(\tilde{\mathbb{T}})$ could be different, and the corresponding loss functions $L(\eta_{\mathbb{T},\mathbb{Q}}, \boldsymbol{w}_{1:T}, \mathbb{Y})$ and $L(\eta_{\tilde{\mathbb{T}},\mathbb{Q}}, \boldsymbol{w}_{1:T}, \mathbb{Y})$ could also be different. This observation suggests that there exists some design $\eta_{\mathbb{T}^*}$ that has a small risk. In the next section, we find such a design when m is correctly specified.

3. Design of Regular Switchback Experiments Under Minimax Rule

The goal of this section is to find the optimal design of regular switchback experiments, that is, to select the

optimal randomization points and the optimal randomization probabilities. Throughout this section, we assume m is known and we set $p = m$.

We formalize our experimental design problem through the minimax framework. The minimax decision rule (Wu 1981, Li 1983, Berger 2013) finds an optimal design of experiment such that the worst-case risk against an adversarial selection of potential outcomes is minimized,

$$\min_{\mathbb{T}\in[T],\,\mathbb{Q}\in\mathcal{Q}} \max_{\mathbb{Y}\in\mathcal{Y}} r(\eta_{\mathbb{T},\mathbb{Q}}, \mathbb{Y}) = \min_{\mathbb{T}\in[T],\,\mathbb{Q}\in\mathcal{Q}} \max_{\mathbb{Y}\in\mathcal{Y}} \sum_{\boldsymbol{w}_{1:T}\in\{0,1\}^T} \eta_{\mathbb{T},\mathbb{Q}}(\boldsymbol{w}_{1:T}) \cdot \left(\widehat{\tau}_p(\boldsymbol{w}_{1:T},\mathbb{Y}) - \tau_p(\mathbb{Y})\right)^2. \tag{6}$$

One compelling reason to adopt the minimax framework, as commented in the seminal work of Wu (1981, p. 1168), is that "the experimenter's information about the model is never perfect. When a model is proposed, there is always the possibility that the 'true' model deviates from the assumed model." Instead of finding the best possible design by imposing a model, we try to derive the best possible design for the worse possible set of potential outcomes.

To overcome minimaxity and to lay out the foundation for inference, we impose an additional assumption on the support of the potential outcome. Because the potential outcomes are unknown but fixed, we assume that their absolute values are bounded from above and that bound is attainable at every time period.

Assumption 3 (Bounded Potential Outcomes). *The potential outcomes are bounded by some constant, that is,* $\exists B > 0, s.t.\ \forall t \in [T],\ \forall \boldsymbol{w} \in \{0,1\}^T, |Y_t(\boldsymbol{w})| \le B$, *or, equivalently,* $\mathbb{Y} \in \mathcal{Y} = [-B, B]^T$.

Assumption 3 is often satisfied because it assumes that the potential outcomes are bounded by the same (possibly a large) constant, (e.g., the ride-matching rate from each experimental period is always a finite quantity) and that the extreme could possibly occur at any point in time (e.g., the maximum ride-matching rate could be observed at any time). In particular, knowledge about the magnitude of B is not required; as we show below, the optimal design does not depend on B.

The reason to make Assumption 3 is twofold. First, for optimization purposes, Assumption 3 reflects the inherent symmetry in the potential outcomes under both treatment and control, which is in the same spirit as the permutation invariance assumption (Wu 1981, Li 1983, Basse et al. 2019b). It is such symmetry that ensures the optimality of fair coin flipping. See Theorem 1. Second, for inferential purposes, Assumption 3 ensures that the variance of the estimator is well behaved, which is commonly assumed in the finite-sample inference literature (Aronow et al. 2017, Chin 2018, Bojinov and Shephard 2019, Li et al. 2020, Han et al. 2021). It is the well-behaved variance that lays the foundation of our limiting distribution of the estimator.

To solve the minimax Problem (6), we start by focusing on the inner maximization part. We characterize the worst-case potential outcomes by identifying two dominating strategies for the adversarial selection of potential outcomes. Denote $\mathbb{Y}^+ = \{Y_t(\mathbf{1}_{m+1}) = Y_t(\mathbf{0}_{m+1}) = B\}_{t\in\{m+1:T\}}$ and $\mathbb{Y}^- = \{Y_t(\mathbf{1}_{m+1}) = Y_t(\mathbf{0}_{m+1}) = -B\}_{t\in\{m+1:T\}}$.

Lemma 1. *Under Assumptions 1–3,* $\mathbb{Y}^+$ *and* $\mathbb{Y}^-$ *are the only two dominating strategies for the adversarial selection of potential outcomes. That is, for any* $\mathbb{T} \subseteq [T]$ *and for any* $\mathbb{Y} \in \mathcal{Y}$,

$$r(\eta_{\mathbb{T},\mathbb{Q}}, \mathbb{Y}^+) \ge r(\eta_{\mathbb{T},\mathbb{Q}}, \mathbb{Y}); \quad r(\eta_{\mathbb{T},\mathbb{Q}}, \mathbb{Y}^-) \ge r(\eta_{\mathbb{T},\mathbb{Q}}, \mathbb{Y}).$$

Moreover, for any $\mathbb{Y} \in \mathcal{Y}$ *such that* $\mathbb{Y} \neq \mathbb{Y}^+, \mathbb{Y} \neq \mathbb{Y}^-$, *the above two inequalities are strict.*

The proof of Lemma 1 can be found in Section EC.3.3.1 in the online appendix. Lemma 1 simplifies the minimax problem in (6), as it allows us to replace $\mathbb{Y}$ by $\mathbb{Y}^* = \mathbb{Y}^+$ or $\mathbb{Y}^* = \mathbb{Y}^-$ and reduce the minimax Problem (6) into a minimization problem

$$\min_{\mathbb{T}\in[T],\,\mathbb{Q}\in\mathcal{Q}} r(\eta_{\mathbb{T},\mathbb{Q}}, \mathbb{Y}^*).$$

Next we solve this minimization problem by first finding the optimal $\mathbb{Q}$ values.

Theorem 1 (Optimality of Fair Coin Flipping). *Under Assumptions 1–3, any optimal design of experiment* $(\mathbb{T}, \mathbb{Q})$ *must satisfy* $q_0 = q_1 = \cdots = q_K = 1/2$.

The proof of Theorem 1 can be found in Section EC.3.4.1 in the online appendix. Theorem 1 suggests that the optimal randomization probabilities should be one-half. So we can restrict our scope to only finding the experiments induced by fair coin flipping and focus on the trade-off behind the number and timing of the randomization points.

The trade-off lies between having too many randomization points (corresponding to large K) and too few randomization points (corresponding to small K). Intuitively, too many decreases the probability of observing consecutive treatments $\mathbf{1}_{m+1}$ or controls $\mathbf{0}_{m+1}$, which, in turn, decreases the amount of useful data. On the other hand, too few decreases the number of independent observations and reduces our ability to produce reliable results. Both of these scenarios reduce our ability to draw valid causal claims. Theorem 2 formalizes this trade-off.

Theorem 2 (Optimal Design). *Under Assumptions 1–3, the optimal solution to the design of regular switchback experiment as we have introduced in (6) is equivalent to the*

optimal solution to the following subset selection problem.

$$\min_{\mathbb{T}\subset[T]} \left\{ 4\sum_{k=0}^{K}(t_{k+1}-t_k)^2 + 8m(t_K-t_1) + 4m^2K - 4m^2 + 4\sum_{k=1}^{K-1}[(m-t_{k+1}+t_k)^+]^2 \right\} \tag{7}$$

In particular, when $m = 0$, *then* $\mathbb{T}^* = \{1,2,3,\ldots,T\}$; *when* $m > 0$, *and if there exists* $n \geq 4 \in \mathbb{N}$, *such that* $T = nm$, *then* $\mathbb{T}^* = \{1, 2m+1, 3m+1, \ldots, (n-2)m+1\}$.

The proof of Theorem 2 is deferred to Section EC.3.6.1 in the online appendix. Theorem 2 presents the optimal design in a class of perfect cases when the time horizon splits into several equal-length epochs;[4] see Table 1 for an example. In practice, we recommend selecting T that satisfies the condition in Theorem 2; see Section 6 for a discussion.

There are two important implications of Theorem 2. First, the optimal randomization frequency depends on the physical duration of the carryover effect, regardless of the granularity of one single experimental period. This observation suggests that practitioners should set each period to be almost as long as the order of the carryover effect, which sheds some light on the selection of granularity when practitioners design the experiment. See Example 7. Second, a special case arises when there are no carryover effects ($m = 0$) or very little carryover effect ($m = 1$); in both cases the optimal designs are almost the same. This observation suggests a layer of robustness when the granularity is set to be the same as the suspected order of the carryover effect; see Example 8.

Example 7 (Two Granularity Levels). In the ride-sharing application, suppose the firm has two options to treat one single time period either as 0.5 hour or one hour; and suppose the carryover effect lasts for two hours. When one single experimental period corresponds to 0.5 hour, the carryover effect lasts for $m = 4$ periods. When one single experimental period corresponds to one hour, the carryover effect lasts for $m = 2$ periods. From Theorem 2, the optimal design exhibits an optimal structure that randomizes once every m periods (except for the first and last epoch, which lasts for $2m$ time periods each). In both cases, the optimal design would randomize once every two hours.

Example 8 (Little Carryover Effect). For example, Theorem 2 suggests that the optimal design when $m = 0$ is $\mathbb{T}^* = \{1,2,3,\ldots,T\}$ and when $m = 1$ is $\mathbb{T}^* = \{1,3,4,\ldots,T-1\}$. This suggests that the minimax optimal design in the absence of a carryover effect is robust to the existence of a short carryover effect.

4. Inference and Testing

After designing and running the experiment, we obtain two time series. The first is the observed assignment path $\boldsymbol{w}^{\text{obs}}_{1:T}$, and the second is the corresponding observed outcomes $\boldsymbol{Y}^{\text{obs}}_{1:T}$. See Figure 3. To draw inference from these data, we propose two methods: an exact randomization based test and a finite population conservative test that establishes asymptotic result.

In Sections 4.1 and 4.2, we assume perfect knowledge of m, that is, $p = m$; we will write τ_m and $\widehat{\tau}_m$ to stand for τ_p and $\widehat{\tau}_p$, respectively. We discuss in Section 4.3 the case when $p \neq m$ and show that our inference methods are still valid. To conclude this section, we provide in Section 4.4 a data-driven procedure to identify a possible value for the carryover effect by running multiple experiments. Such a procedure relaxes Assumption 2 and is of great practical relevance.

4.1. Exact Inference

We propose an exact nonparametric test for the sharp null of no effect at every time point (Fisher et al. 1937, Rubin 1980, Bojinov and Shephard 2019):

$$H_0 : Y_t(\boldsymbol{w}_{t-m:t}) - Y_t(\boldsymbol{w}'_{t-m:t}) = 0 \quad \text{for all } \boldsymbol{w}_{t-m:t}, \boldsymbol{w}_{t-m} : t', \quad t \in \{m+1 : T\}. \tag{8}$$

The sharp null hypothesis implies that $Y_t(\boldsymbol{w}^{\text{obs}}_{t-m:t}) = Y_t(\boldsymbol{w}_{t-m:t})$ for all $\boldsymbol{w}_{t-m:t} \in \{0,1\}^t$. That is, regardless of the assignment path $\boldsymbol{w}_{t-m:t}$, we would have observed the same outcomes.

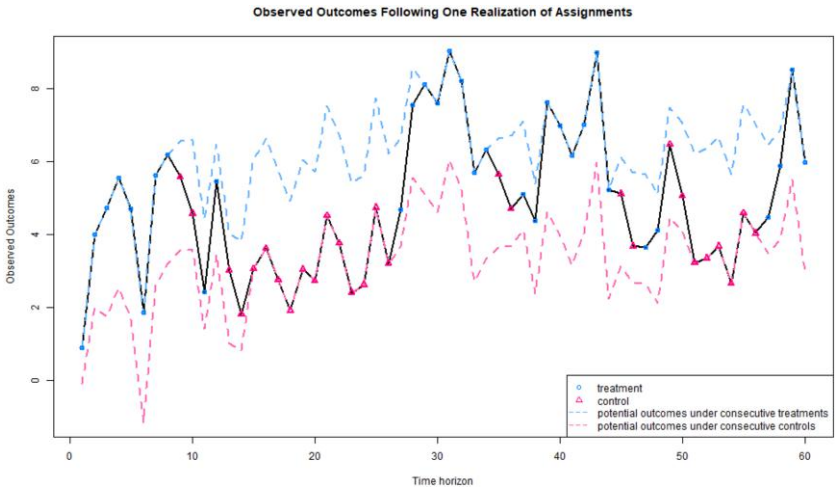
We can conduct exact tests by using the known assignment mechanism to simulate new assignment paths; see Algorithm 1 for details. The test depends on the observation that, under the sharp null hypothesis of no treatment effect (8), any assignment path $\boldsymbol{w}^{[i]}_{1:T}$ leads to the same observed outcomes. In particular, in Step 3, we assume the observed outcomes remain unchanged. Thus, all treatment paths lead to the same observed outcomes $Y^{\text{obs}}_{m+1:T}$. To obtain a confidence interval, we propose inverting a sequence of exact hypothesis tests to identify the region outside of which (8) is violated at the prespecified nominal level (Imbens and Rubin 2015, chapter 5). In practice, obtaining confidence intervals through this approach is somewhat challenging; instead, we refer the reader to the subsequent section that provides a less computationally intensive approach.

Table 1. An Example of the Optimal Design $\mathbb{T}^* = \{1,5,7,9\}$ When $T = 12$ and $p = m = 2$

	1	2	3	4	5	6	7	8	9	10	11	12
$\mathbb{T}^*$	✓	–	–	–	✓	–	✓	–	✓	–	–	–

Note. Each checkmark beneath a time period t indicates that t is a randomization point.

Figure 3. (Color online) Illustrator of the Observed Assignment Path $w^{\text{obs}}_{1:T}$ (Blue and Red Dots) and the Observed Outcomes $Y^{\text{obs}}_{p+1:T}$ (Black Curve)



Note. The dashed lines are the potential outcomes under consecutive treatments/controls.

Algorithm 1 (Algorithm for Performing a Sharp-Null Hypothesis Test)

Require: Fix I, total number of samples drawn.

1: **for** i in 1 : I **do**

2: Sample a new assignment path $\boldsymbol{w}^{[i]}_{1:T}$ according to the assignment mechanism.

3: Hold $Y^{\text{obs}}_{p+1:T}$ unchanged. Compute $\widehat{\tau}^{[i]}$ according to (4),

$$\widehat{\tau}^{[i]} = \frac{1}{T-m}\sum_{t=m+1}^{T}\left\{Y^{\text{obs}}_t \frac{\mathbb{1}\{\boldsymbol{w}^{[i]}_{t-m:t} = \mathbf{1}_{m+1}\}}{\Pr(\boldsymbol{W}_{t-m:t} = \mathbf{1}_{m+1})} - Y^{\text{obs}}_t \frac{\mathbb{1}\{\boldsymbol{w}^{[i]}_{t-m:t} = \mathbf{0}_{m+1}\}}{\Pr(\boldsymbol{W}_{t-m:t} = \mathbf{0}_{m+1})}\right\}.$$

4: **end for**

5: Compute $\widehat{p}_{\mathsf{F}} = I^{-1}\sum_{i=1}^{I}\mathbb{1}\{|\widehat{\tau}^{[i]}| > |\widehat{\tau}|\}$

6: **return** $\widehat{p}_{\mathsf{F}}$, the estimated p-value. For large I, this is exact.

4.2. Asymptotic Inference

We now introduce a conservative test for the null of no average treatment effect:

$$H_0 : \tau_m = \frac{1}{T-m}\sum_{t=m+1}^{T}[Y_t(\mathbf{1}_{m+1}) - Y_t(\mathbf{0}_{m+1})] = 0. \tag{9}$$

To test such a null, we derive a finite population central limit theorem to approximate the distribution of the Horvitz-Thompson estimator.

Assume $n = T/m \geq 4$ is an integer, then under the optimal design as shown in Theorems 1 and 2, the assignment path is determined by the realizations at $W_1, W_{2m+1}, \ldots, W_{(n-2)m+1}$. To make the dependence on randomization clear, we introduce the following notations. For any $k \in \{0,1,\ldots,n-2\}$, let $\bar{Y}_k(\mathbf{1}_{m+1}) = \sum_{t=(k+1)m+1}^{(k+2)m} Y_t(\mathbf{1}_{m+1})$ and $\bar{Y}_k(\mathbf{0}_{m+1}) = \sum_{t=(k+1)m+1}^{(k+2)m} Y_t(\mathbf{0}_{m+1})$. Moreover, for any $k \in \{0,1,\ldots,n-2\}$, let $\bar{Y}^{\text{obs}}_k = \sum_{t=(k+1)m+1}^{(k+2)m} Y^{\text{obs}}_t$ be the sum of the observed outcomes.

Lemma 2 (Variance of the Horvitz-Thompson Estimator Under the Optimal Design). *Under Assumptions 1–3 and under the optimal design as shown in Theorems 1 and 2, if $n = T/m \geq 4$ is an integer, then the variance of the Horvitz-Thompson estimator,* $\mathsf{Var}(\widehat{\tau}_m)$, *is*

$$\begin{aligned}\mathsf{Var}(\widehat{\tau}_m) = \frac{1}{(T-m)^2}\Bigg\{&\bar{Y}_0(\mathbf{1}_{m+1})^2 + \bar{Y}_0(\mathbf{0}_{m+1})^2 + 2\bar{Y}_0(\mathbf{1}_{m+1})\bar{Y}_0(\mathbf{0}_{m+1}) \\ &+ \sum_{k=1}^{n-3}\left[3\bar{Y}_k(\mathbf{1}_{m+1})^2 + 3\bar{Y}_k(\mathbf{0}_{m+1})^2 + 2\bar{Y}_k(\mathbf{1}_{m+1})\bar{Y}_k(\mathbf{0}_{m+1})\right] \\ &+ \bar{Y}_{n-2}(\mathbf{1}_{m+1})^2 + \bar{Y}_{n-2}(\mathbf{0}_{m+1})^2 + 2\bar{Y}_{n-2}(\mathbf{1}_{m+1})\bar{Y}_{n-2}(\mathbf{0}_{m+1}) \\ &+ \sum_{k=0}^{n-3} 2\left[\bar{Y}_k(\mathbf{1}_{m+1}) + \bar{Y}_k(\mathbf{0}_{m+1})\right]\cdot\left[\bar{Y}_{k+1}(\mathbf{1}_{m+1}) + \bar{Y}_{k+1}(\mathbf{0}_{m+1})\right]\Bigg\}.\end{aligned} \tag{10}$$

Lemma 2 provides the variance of the Horvitz-Thompson estimator under the optimal design. Because we never observe all the potential outcomes, most of the cross-product terms in (10) cannot be directly estimated. Instead, we provide the following upper bound to (10) and propose an unbiased estimator.

Corollary 1. *Under the conditions in Lemma 2, there exists an upper bound for the variance of the Horvitz-Thompson*

estimator, $\mathsf{Var}(\widehat{\tau}_m) \le \mathsf{Var}^{\mathsf{U}}(\widehat{\tau}_m)$*, which can be estimated by* $\widehat{\sigma}^2_{\mathsf{U}}$*, defined as*

$$\widehat{\sigma}^2_{\mathsf{U}} = \frac{1}{(T-m)^2}\left\{8(\bar{Y}_0^{\mathsf{obs}})^2 + \sum_{k=1}^{n-3} 32(\bar{Y}_k^{\mathsf{obs}})^2\, \mathbb{1}\{W_{km+1} = W_{(k+1)m+1}\} + 8(\bar{Y}_{n-2}^{\mathsf{obs}})^2\right\}.$$

Moreover, $\widehat{\sigma}^2_{\mathsf{U}}$ *is unbiased, that is.,* $\mathbb{E}[\widehat{\sigma}^2_{\mathsf{U}}] = \mathsf{Var}^{\mathsf{U}}(\widehat{\tau}_m)$.

Corollary 1 provides the foundation to make conservative inference. We make the following technical assumption for the asymptotic normal distribution to hold.

Assumption 4 (Nonnegligible Variance). *Assume that the randomization distribution has a nonnegligible variance, that is,*

$$\mathsf{Var}(\widehat{\tau}_m) \ge \Omega(n^{-1}). \tag{11}$$

In particular, one sufficient condition for (11) *is to assume that all the potential outcomes are positive, that is, there exists some constant* $b > 0$*, such that* $\forall t \in [T]$, $\forall \boldsymbol{w}_{1:T} \in \{0,1\}^T$, $Y_t(\boldsymbol{w}_{1:T}) \ge b$.

Intuitively, the key to most central limit theorems is that all the variables roughly have variances of the same order. In other words, there cannot be a small number of variables that compromise the majority of the variance. Because under Assumption 3 the potential outcomes are bounded, each variable contributes to the total variance of order $O(n^{-2})$. Assumption 4 suggests that the total variance is large enough, such that it cannot come from only a few of the time periods.

Theorem 3 (Asymptotic Normality). *Let* m *be fixed. For any* $n \ge 4 \in \mathbb{N}$*, define an* n*-replica experiment such that there are* $T = nm$ *time periods. We take the optimal design as in Theorem 2 whose randomization points are at* $\mathbb{T}^* = \{1, 2m+1, 3m+1, \ldots, (n-2)m+1\}$*. Under Assumptions 1–2, and under Assumption 4, the limiting distribution of the Horvitz-Thompson estimator in the* n*-replica experiment has an asymptotic normal distribution. That is, let* $\mathsf{Var}(\widehat{\tau}_m)$ *be defined in Lemma 2. As* $n \to +\infty$,

$$\frac{\widehat{\tau}_m - \tau_m}{\sqrt{\mathsf{Var}(\widehat{\tau}_m)}} \xrightarrow{D} \mathcal{N}(0,1).$$

Theorem 3 is in the spirit of the finite population central limit theorems as in Li and Ding (2017), Aronow et al. (2017), Chin (2018), Bojinov et al. (2021), and Han et al. (2021). Note that Theorem 3 does not require $\mathsf{Var}(\widehat{\tau}_m)$ to converge as $n \to +\infty$.

To conduct inference, we replace $\mathsf{Var}(\widehat{\tau}_m)$ by $\widehat{\sigma}^2_{\mathsf{U}}$ as provided in Corollary 1. Define the test statistic to be $z = |\widehat{\tau}_m| / \sqrt{\widehat{\sigma}^2_{\mathsf{U}}}$. When the alternative hypothesis is two sided, the estimated p-value is given by $\widehat{p}_{\mathsf{N}} = 2 - 2\Phi(z)$, where Φ is the cumulative distribution function (CDF) of a standard normal distribution.

The proofs of Lemma 2, Corollary 1, and Theorem 3 are deferred to Sections EC.4.2, EC.4.3, and EC4.4 in the online appendix, respectively.

4.3. Inference Under Misspecified *m*

Up to now, we assumed that we knew the order of the carryover effect m and set $p = m$. In practice, we may not know the exact value of the carryover effect, and we have to select p either based on domain knowledge or the procedure we recommend in Section 4.4. In this section, we consider what happens when $p \ne m$ and show that the estimation and inference are still valid and meaningful, although the design from Theorem 2 is no longer optimal. Below we distinguish two cases: $p > m$ and $p < m$.

When $p > m$, because of Assumption 2, $Y_t(\mathbf{1}_{p+1}) = Y_t(\mathbf{1}_{m+1})$, $\forall t \in \{p+1 : T\}$, and the lag-$p$ causal effect is essentially the lag-m causal effect. So all the estimation and inference results still hold.

However, when $p < m$, the Horvitz-Thompson estimator (4) will be biased for the causal estimand. See Section EC.4.5 in the online appendix for more discussions. When $p < m$, the exact inference procedure as in Section 4.1 remains valid. For the asymptotic inference procedure, a similar result to Theorem 3 still holds when m is misspecified, as we state in Corollary 2. The only difference is that when $p < m$, the asymptotic normal distribution will not be centered around the causal estimand as we defined in (1) but some quantity that we will discuss in Section EC.4.5. The proof is deferred to Section EC.4.7 in the online appendix.

Corollary 2 (Asymptotic Normality When m is Misspecified). *For any* $n \ge 4 \in \mathbb{N}$*, define an* n*-replica experiment such that there are* $T = np$ *time periods. Take the optimal design as in Theorem 2 whose randomization points are at* $\mathbb{T}^* = \{1, 2p+1, 3p+1, \ldots, (n-2)p+1\}$*. We have the following two observations.*

i. *When* $p > m$*, under Assumptions 1–2, the variance of the Horvitz-Thompson estimator,* $\mathsf{Var}(\widehat{\tau}_p)$*, is explicitly given by* (10).

ii. *Furthermore, no matter if* $p > m$ *or* $p < m$*, under Assumptions 1–3 and assume* $\mathsf{Var}(\widehat{\tau}_p) \ge \Omega(n^{-1})$*, the limiting distribution of the Horvitz-Thompson estimator in the* n*-replica experiment has an asymptotic normal distribution. That is, as* $n \to +\infty$,

$$\frac{\widehat{\tau}_p - \mathbb{E}[\widehat{\tau}_p]}{\sqrt{\mathsf{Var}(\widehat{\tau}_p)}} \xrightarrow{D} \mathcal{N}(0,1).$$

Corollary 2, together with Theorem 3, is the key to identification of m, the order of the carryover effect. In Section 4.4, we provide a procedure to identify m.

4.4. Identifying the Order of the Carryover Effect

Using Theorem 3 and Corollary 2, we can define a hypothesis testing procedure, which, combined with a searching

method, yields an estimate of the order of the carryover effect.

To build intuition, suppose we have access to two comparable experimental units. The two experimental units could be two separate units or two nonoverlapping time epochs on one experimental unit such that the two epochs are far enough such that the carryover effect from one does not affect the outcomes of the other. Suppose on the first experimental unit we design an optimal experiment under $p = p_1$, and on the second unit we use $p = p_2$; without loss of generality, let $p_1 < p_2$.

After running the experiment and collecting the results, consider the following two statistics. For the first unit, we calculate $\widehat{\tau}_{p_1}$, the sampling average, and $\widehat{\sigma}^2_{p_1}$, the conservative sampling variance as suggested by Corollary 1. For the second unit, we calculate $\widehat{\tau}_{p_2}$ and $\widehat{\sigma}^2_{p_2}$.

Define a procedure that tests the following null hypothesis:

$$H_0 : m \leq p_1. \tag{12}$$

Under the null Hypothesis (12), ; so both $\widehat{\tau}_{p_1}$ and $\widehat{\tau}_{p_2}$ are unbiased estimators of τ_m. Furthermore, given that the two estimators both conform asymptotic normal distributions and that the two experimental units are independent, the difference between the two estimators should be an asymptotic normal distribution centered around zero, that is,

$$(\widehat{\tau}_{p_1} - \widehat{\tau}_{p_2}) / \sqrt{\mathsf{Var}(\tau_{p_1}) + \mathsf{Var}(\tau_{p_2})} \xrightarrow{D} \mathcal{N}(0,1).$$

To test the null Hypothesis (12), define the test statistic to be

$$z = \left|\widehat{\tau}_{p_1} - \widehat{\tau}_{p_2}\right| / \sqrt{\widehat{\sigma}^2_{p_1} + \widehat{\sigma}^2_{p_2}}.$$

The estimated p-value is given by $\widehat{p} = 2 - 2\Phi(z)$, where Φ is the CDF of a standard normal distribution.

The above procedure enables us to test the null Hypothesis (12). We can combine such a procedure with any searching method to identify m.

5. Simulation Study

There are five goals for this simulation study. First, show that the optimal design in Theorem 2 has the smallest risk compared against two benchmarks. There are two dimensions for our comparison: the worst-case risk and the risk under a specific outcome model. Second, verify the asymptotic normal distribution under a nonasymptotic setup and study the quality of the upper bound proposed in Corollary 1. Third, understand the rejection rate and its dependence on the length of time horizon. Fourth, study the performance of the optimal design under a misspecified m and compare the difference of the two inference methods proposed in Section 4. Fifth, study the performance of the hypothesis testing procedure as proposed in Section 4.4, which identifies m the length of the carryover effect.

We start with a simple linear additive carryover effect model, which originates from Hedayat et al. (1978), Oman and Seiden (1988), and Jones and Kenward (2014).

$$Y_t(\boldsymbol{w}_{1:t}) = \mu + \alpha_t + \delta^{(1)} w_t + \delta^{(2)} w_{t-1} + \cdots + \delta^{(t)} w_1 + \epsilon_t, \tag{13}$$

where μ is a fixed effect; α_t is a fixed effect associated to period t; $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(t)}$ are nonstochastic coefficients; $w_t, w_{t-1}, \ldots, w_1$ are the treatment indicators; and ϵ_t is the random noise in period t. We will run many simulations based on this model. For a more detailed discussion of the flexibility of the potential outcome framework, see Section EC.5.1 in the online appendix.

5.1. Comparison of the Risk Functions for Different Designs

5.1.1. Simulation Setup. We consider two setups. The first setup is for the worst-case risk. We consider $T = 120$, $p = m = 2$, where m is correctly identified, and $Y_t(\mathbf{1}_3) = Y_t(\mathbf{0}_3) = 10$. We compare three different designs of switchback experiments. The first one is our proposed optimal design as in Theorem 2, such that $\mathbb{T}^* = \{1, 5, 7, \ldots, 117\}$. The second one is the most common and naive switchback experiment, which independently assigns treatment/control in every period with half-half probability. It is parameterized by $\mathbb{T}^{\mathsf{H1}} = \{1, 2, 3, \ldots, 120\}$. The third one is the "intuitive" experiment discussed in Table 1, which divides the time horizon into several epochs each with length $m + 1 = 3$. It is parameterized by $\mathbb{T}^{\mathsf{H2}} = \{1, 4, 7, \ldots, 118\}$.

Second, we run simulations based on the outcome model as in (13). Similar to the first setup, we consider again $T = 120, p = m = 2$ where m is correctly identified. For the outcome model, we consider $\mu = 0$, $\alpha_t = \log(t)$, and $\epsilon_t \sim N(0,1)$ are independent and identically distributed (i.i.d.) standard normal distributions. For any $t > 3$, let $\delta^{(t)} = 0$. We will vary the values of $\delta^{(1)}, \delta^{(2)}, \delta^{(3)} \in \{1, 2\}$ and conduct experiments under $2^3 = 8$ different scenarios. Again we compare the same three different designs of switchback experiments: $\mathbb{T}^* = \{1, 5, 7, \ldots, 117\}$, $\mathbb{T}^{\mathsf{H1}} = \{1, 2, 3, \ldots, 120\}$, and $\mathbb{T}^{\mathsf{H2}} = \{1, 4, 7, \ldots, 118\}$.

We simulate one assignment path at a time and conduct an experiment following this assignment path. Because the outcome model is prescribed, we can calculate both the causal estimand and the observed outcomes (along the simulated assignment path). Then we calculate the Horvitz-Thompson estimator based on the simulated assignment path and the simulated observed outcomes. With both the estimand and estimator, we can calculate the loss function. We repeat the above procedure enough (100,000) times to obtain an accurate approximation of the risk function.

5.1.2. Simulation Results. First, we calculate the worst-case risk functions via simulations. Notice that, when $p = m = 2$, we could explicitly calculate the worst-case risk functions under the three different designs of switchback experiments $\mathbb{T}^*$, $\mathbb{T}^{\mathsf{H1}}$, and $\mathbb{T}^{\mathsf{H2}}$. Even though we can explicitly calculate them via the following expression (see Lemma EC.9 in the online appendix for details),

$$\frac{B^2}{(T-m)^2}\left\{4\sum_{k=1}^{K+1}(t_k-t_{k-1})^2+8m(t_K-t_1)+4m^2K-4m^2+4\sum_{k=2}^{K}[(m-t_k+t_{k-1})^+]^2\right\}, \tag{14}$$

we still use the simulation to confirm this result. See Table 2 for our simulation results.

The causal effect is $\tau_2 = 0$ because $Y_t(\mathbf{1}_3) = Y_t(\mathbf{0}_3) = 10$. The simulated estimator is $\mathbb{E}[\widehat{\tau}_2^*] = -0.0291$ for our proposed optimal design and $\mathbb{E}[\widehat{\tau}_2^{\mathsf{H1}}] = 0.0104$ and $\mathbb{E}[\widehat{\tau}_2^{\mathsf{H2}}] = -0.0478$ for the two benchmarks, respectively. The risk function is $r(\eta_{\mathbb{T}^*}) = 26.78$ for our proposed optimal design and $r(\eta_{\mathbb{T}^{\mathsf{H1}}}) = 33.67$ and $r(\eta_{\mathbb{T}^{\mathsf{H1}}}) = 27.85$ for the two benchmarks, respectively. Such simulation results suggest that our proposed optimal design has the smallest risk, under the worst-case outcome model. In the last three columns are the risk functions of the three designs, all suggested by Expression (14). The risk functions calculated from theory take values that are very close to the risk functions calculated from Expression (14), which verifies our theory.

Second, we calculate the risk functions based on the outcome model in (13). See Table 3. As we vary the values of $\delta^{(1)}$, $\delta^{(2)}$, and $\delta^{(3)}$, the average lag-2 causal effect is being changed. All three estimators are able to reflect the change as the estimand changes. The risk function can be simulated, and we see that the risk function associated with the first benchmark $\mathbb{T}^{\mathsf{H1}}$ is 28%−32% larger than the optimal design; the second benchmark $\mathbb{T}^{\mathsf{H2}}$ is 1%−2% larger. Such simulation results suggest again that our proposed optimal design has the smallest risk. Moreover, as $r(\eta_{\mathbb{T}^{\mathsf{H2}}})$ is close to $r(\eta_{\mathbb{T}^*})$ and both are much smaller than $r(\eta_{\mathbb{T}^{\mathsf{H1}}})$, our results suggest that when m is unknown, it is better to select p to be slightly larger than the true m as opposed to significantly smaller.

As the magnitude of treatment effects increase, the associated risk functions also increase. The relative difference between risk functions of $r(\eta_{\mathbb{T}^{\mathsf{H1}}})$ and $r(\eta_{\mathbb{T}^*})$ increases, whereas the relative difference between $r(\eta_{\mathbb{T}^{\mathsf{H1}}})$ and $r(\eta_{\mathbb{T}^*})$ decreases. This coincides with the intuitions discussed in Section 3.

5.2. Asymptotic Normality

5.2.1. Simulation Setup. We run simulations based on the outcome model in (13), with $T = 120$ and $m = 2$. We will consider three cases: (i) m is correctly specified, so $p = 2$; (ii) $p = 3$, and we estimate lag-3 causal estimand as in (1); (iii) $p = 1$, and we pretend as if we estimated the lag-1 causal estimand. However, as the lag-1 causal estimand is not well defined, we instead estimate a different quantity, which we refer to as the "m-misspecified lag-p causal estimand" (see details and definition in (EC.10) in the online appendix).

For the outcome model, we consider $\mu = 0$, $\alpha_t = \log(t)$, and $\epsilon_t \sim N(0,1)$ are i.i.d. standard normal distributions. For any $t > 3$, let $\delta^{(t)} = 0$. For simplicity, let $\delta^{(1)} = \delta^{(2)} = \delta^{(3)} = \delta$. We vary $\delta \in \{1,2,3\}$ and conduct experiments under three different scenarios. We simulate one assignment path at a time and conduct experiments following this assignment path. Because the outcome model is prescribed, we calculate the observed outcomes based on the simulated assignment path. Then we calculate the Horvitz-Thompson estimator, and the conservative estimator of the randomization variance (Corollary 1), based on the simulated assignment path and the simulated observed outcomes. On the other hand, the lag-p causal estimand is easy to calculate once the outcome model is prescribed. Yet the m-misspecified lag-p causal estimand has to be calculated in conjunction with the simulated assignment path. By repeating the above procedure enough (100,000) times, we obtain a distribution of the estimator.

5.2.2. Simulation Results. In Figure 4, the dotted dark blue line is the probability density function of the standard normal distribution. The pink histogram corresponds to the distribution induced by $\frac{\widehat{\tau}_p-\tau_p}{\sqrt{\mathsf{Var}(\widehat{\tau}_p)}}$, which is the estimator (after recentering at zero) normalized by the square root of the true randomization variance.[5] Such a distribution, as suggested by Theorem 3, converges to a standard normal distribution when T is large. Comparing to the dotted dark blue line, Figure 4 suggests that Theorem 3 approximately holds for moderate values of T. The light blue histogram corresponds to the distribution induced by $\frac{\widehat{\tau}_p-\tau_p}{\sqrt{\mathbb{E}[\widehat{\sigma}_U^2]}}$, which is the estimator (after recentering at zero) normalized by the expectation of the conservative upper bound of the

Table 2. Simulation Results for the Worst-Case Risk Function

τ_2	$\mathbb{E}[\widehat{\tau}_2^*]$	$\mathbb{E}[\widehat{\tau}_2^{\mathsf{H1}}]$	$\mathbb{E}[\widehat{\tau}_2^{\mathsf{H2}}]$	$r(\eta_{\mathbb{T}^*})$	$r(\eta_{\mathbb{T}^{\mathsf{H1}}})$	$r(\eta_{\mathbb{T}^{\mathsf{H2}}})$	$\tilde{r}(\eta_{\mathbb{T}^*})$	$\tilde{r}(\eta_{\mathbb{T}^{\mathsf{H1}}})$	$\tilde{r}(\eta_{\mathbb{T}^{\mathsf{H2}}})$
0	0.0250	0.0200	0.0059	26.78	33.67	27.85	26.67	33.96	27.81

Note. The optimal design $\mathbb{T}^*$ as suggested by Theorem 2 yields the smallest risk, both in theory and confirmed by simulations.

Table 3. Simulation Results for the Risk Function Based on the Outcome Model in (13)

$\delta^{(1)}$	$\delta^{(2)}$	$\delta^{(3)}$	τ_2	$\mathbb{E}[\widehat{\tau}_2^*]$	$\mathbb{E}[\widehat{\tau}_2^{\mathsf{H1}}]$	$\mathbb{E}[\widehat{\tau}_2^{\mathsf{H2}}]$	$r(\eta_{\mathbb{T}^*})$	$r(\eta_{\mathbb{T}^{\mathsf{H1}}})$	$r(\eta_{\mathbb{T}^{\mathsf{H2}}})$
1	1	1	3	3.016	3.012	3.002	7.96	10.22	8.11
1	1	2	4	4.018	4.013	4.002	9.57	12.39	9.74
1	2	1	4	4.018	4.013	4.002	9.57	12.39	9.74
2	1	1	4	4.018	4.013	4.002	9.57	12.39	9.74
1	2	2	5	5.020	5.015	5.003	11.34	14.81	11.52
2	1	2	5	5.020	5.015	5.003	11.34	14.81	11.52
2	2	1	5	5.020	5.015	5.003	11.34	14.81	11.52
2	2	2	6	6.022	6.016	6.003	13.28	17.48	13.47

Notes. For each row, the random seed that generates the simulation setup is fixed. The optimal design $\mathbb{T}^*$ as suggested in Theorem 2, though solved from a minimax program, still yields the smallest risk for the outcome model in (13). A few rows are redundant because our switchback experiment, combining with the causal estimand (1), is only able to measure the total additive treatment effect. We cannot distinguish the source of the additive treatment effects; that is, we are unable to distinguish $\delta^{(1)}$, $\delta^{(2)}$, and $\delta^{(3)}$.

randomization variance. Because we replace the true variance by the conservative upper bound, the shape of the distribution is more concentrated around zero, as we see from the "taller" histogram. The red vertical line is the expected value of the randomization distribution for the pink histogram. The cases of $\delta = 1$ and $\delta = 2$ are similar, and the cases of overestimated m and underestimated m are also similar. We discuss them in Section EC.5.3 in the online appendix.

For all the nine cases ($p \in \{1,2,3\}$ and $\delta \in \{1,2,3\}$), see Table 4 for the expected values and the variances of the randomization distributions as well as the conservative estimator of the randomization variances. Note that the three cases all have the same underlying outcome model. It is the different knowledge of m that leads to three different designs of experiments.

From Table 4, we make the following two observations. *(i) Unbiasedness of the Horvitz-Thompson estimator.* When m is correctly specified, $\mathbb{R}[\widehat{\tau}_p]$ is very close to τ_p, verifying the unbiasedness of the estimator. When $m = 2, p = 3$, the estimand remains unchanged and the estimator remains unbiased. But the variance of the estimator is larger. When $m = 2, p = 1$, the estimand is the

Table 4. Simulation Results for the Randomization Distribution

		τ_p	$\tau_p^{[m]}$	$\mathbb{E}[\widehat{\tau}_p]$	$\mathsf{Var}(\widehat{\tau}_p)$	$\mathbb{E}[\widehat{\sigma}_{\mathsf{U}}^2]$
$m = 2, p = 2$	$\delta = 1$	3	–	3.016	7.96	8.48
	$\delta = 2$	6	–	6.022	13.28	15.16
	$\delta = 3$	9	–	9.028	20.10	24.25
$m = 2, p = 3$	$\delta = 1$	3	–	3.006	11.92	12.67
	$\delta = 2$	6	–	6.009	19.89	22.70
	$\delta = 3$	9	–	9.012	30.10	36.32
$m = 2, p = 1$	$\delta = 1$	–	2	2.016	4.00	4.13
	$\delta = 2$	–	4	4.026	6.69	7.06
	$\delta = 3$	–	6	6.037	10.14	10.92

Notes. The randomization distribution is unbiased in all nine cases (when $p < m$ it is unbiased for the m-misspecified average lag-1 causal effect). The conservative estimation of the variance upper bound from Corollary 1 is close to the true variance.

m-misspecified estimand and the estimator is unbiased for this m-misspecified estimand. *(ii) Quality of Corollary 1 and 2.* As we increase δ, the variances of the randomization distributions also increase. The conservative estimators of the randomization variances are very close to the true variances, which suggests that Corollary 1 and 2 approximate the true variances quite well.

5.2.3. Robustness Check. In this section, we run simulations under almost the same setup as introduced in Section 5.2.1, with the only difference being that we select each ϵ_t to be an i.i.d. student's t-distribution with one degree of freedom. The purpose of this section is to verify our theory when ϵ_t are drawn from heavy tailed distributions.

When $m = 2, p = 2, \delta = 1$, as we can see from Figure 5, the randomization distribution is significantly different from a standard normal distribution. This is because $T = 120$ is too small. Alternatively, we increase $T = 1{,}200$ to see that the randomization distribution behaves like a normal distribution; see Figure 6. In other words, when ϵ_t noises are heavy tailed, our Theorem 3 has a slower convergence rate to a normal distribution. We conduct extensive simulation study under other parameters, as we will show in Section EC.5.3 in the online appendix.

Figure 4. (Color online) Approximate Normality of the Randomization Distribution When $m = 2, p = 2, \delta = 3$

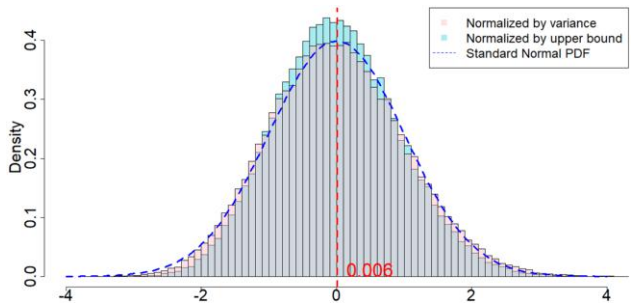


Figure 5. (Color online) Randomization Distribution When Random Noises Are Student's t-Distributions and When $m = 2$, $p = 2, \delta = 1, T = 120$

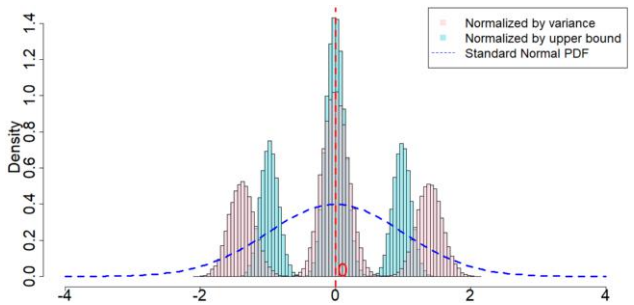
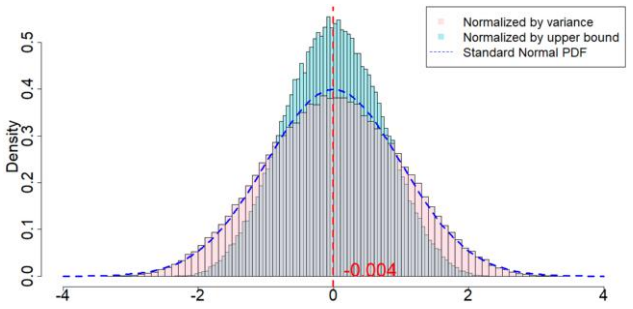


Figure 6. (Color online) Randomization Distribution When Random Noises Are Student's t-Distributions and When $m = 2$, $p = 2, \delta = 1, T = 1{,}200$



5.3. Rejection Rates

5.3.1. Simulation Setup. We run simulations based on the outcome model as in (13). We vary $T \in \{120, 240, \ldots, 1{,}200\}$. We consider $p = m = 2$ where m is correctly specified. Similar to Section 5.2, we consider the same parameterization and conduct experiments under three different scenarios $\delta \in \{1, 2, 3\}$.

We simulate one assignment path at a time, and conduct experiments following this assignment path. We first calculate the observed outcomes and the Horvitz-Thompson estimator. Then we conduct the two inference methods as proposed in Section 4, and obtain two estimated p-values. For the asymptotic inference method, we plug in $\widehat{\sigma}^2_{\mathrm{U}}$, the conservative upper bound of the variance. We reject the corresponding null hypothesis when the p-value is smaller than 0.1. (In Section EC.5.4 in the online appendix, we run additional simulations by replacing such 0.1 threshold by 0.05 and 0.01.) By repeating the above procedure enough (in this simulation, 1000) times, we obtain the frequency of a null hypothesis being rejected, which we refer to as the rejection rate.

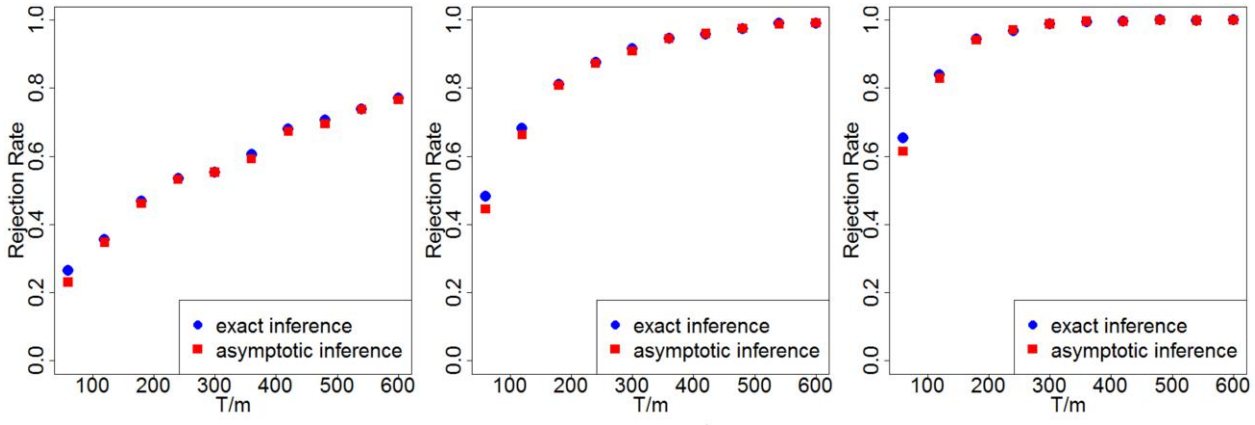
5.3.2. Simulation Results. We calculate the rejection rates via simulations and then plot Figure 7. The blue dots are rejection rates under exact inference; the red dots are under asymptotic inference. In all the simulations, $\delta \neq 0, \tau_p \neq 0$. So, ideally, we would wish to reject both the Fisher's null hypothesis (8) and the Neyman's null hypothesis (9).

From Figure 7, we make the following three observations. *(i) Dependence on T/m.* The rejection rates increase as the length of the horizon increases—more specifically, as T/m the total number of epochs increases. In practice, when firms have the capability to choose the length of T, they can refer to Figure 7 to choose T properly. Also see the discussion in Section 6. *(ii) Between two inference methods.* In all three cases, the rejection rate from testing a sharp null hypothesis (8) is slightly higher than that from testing the Neyman's null (9). This coincides with our intuition that a sharp null is more likely to be rejected. We discuss this in Section 5.5.2 together with the associated p-values. *(iii) Dependence on the signal-to-noise ratio.* The rejection rates all increase as δ increases from one to three (while holding the noise from the model fixed). This suggests that when the treatment effect is relatively larger, we do not require a long experimental horizon to achieve a desired rejection rate.

5.4. Comparison of the Type I and Type II Errors for Different Designs

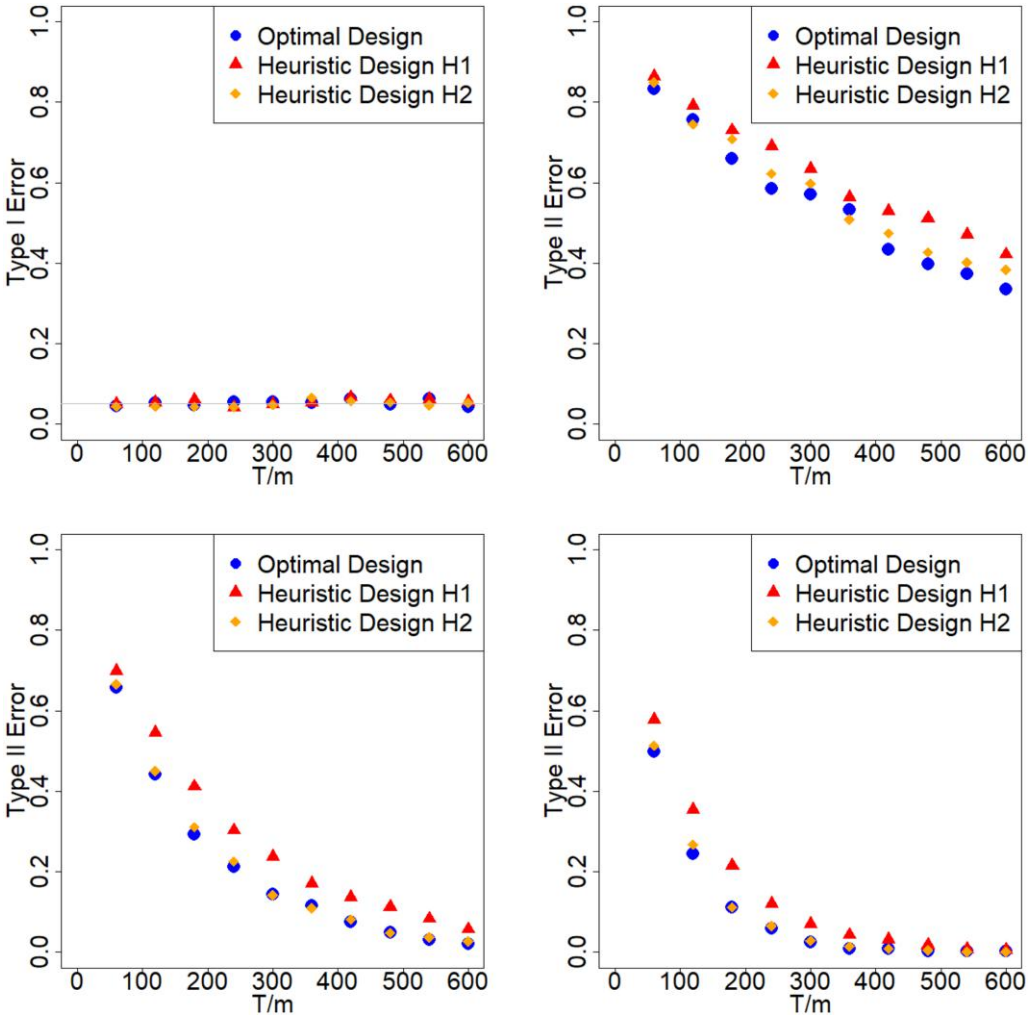
5.4.1. Simulation Setup. We run simulations based on the outcome model as in (13). We vary $T \in \{120, 240, \ldots, 1{,}200\}$. We consider $p = m = 2$ where m is correctly specified. Similar to Section 5.2, we consider the same parameterization and conduct experiments under three different scenarios $\delta \in \{1, 2, 3\}$. We compare three designs of experiments as described in Section 5.1: the optimal design $\mathbb{T}^* = \{1, 5, 7, \ldots, 117\}$, which we refer to as *Optimal Design* as in Figure 8; the most commonly

Figure 7. (Color online) Rejection Rates and Their Dependence on T/m



Note. Left: $\delta = 1$; middle: $\delta = 2$; right: $\delta = 3$.

Figure 8. (Color online) Type I and Type II Errors



Note. Upper left: $\delta = 0$, type I error; upper right: $\delta = 1$, type II error; bottom left: $\delta = 2$, type II error; bottom right: $\delta = 3$, type II error.

adopted heuristic $\mathbb{T}^{H1} = \{1, 2, 3, \ldots, 120\}$, which we refer to as *Heuristic Design H1*; and the so-called intuitive design $\mathbb{T}^{H2} = \{1, 4, 7, \ldots, 118\}$, which we refer to as *Heuristic Design H2*.

In this simulation, we first calculate the frequency of rejecting the Fisher's null hypothesis as in (8) out of a total of 1000 repetitions. And then we use the frequency to calculate the type I and type II errors. Type I error is the probability of rejecting the null hypothesis when there is no treatment effect, which we simulate as the frequency of rejection using $\delta = 0$ when there is no treatment effect. Type II error is the probability of not rejecting the null hypothesis when there is a treatment effect, which we simulate as one minus the frequency of rejection using $\delta \in \{1, 2, 3\}$ when there is a nonnegligible treatment effect.

5.4.2. Simulation Results. The simulation results are summarized in Figure 8. The blue dots are the type I and type II errors of the optimal design; the red dots are the type I and type II errors of the heuristic design *H*1; the yellow dots are the type I and type II errors of the heuristic design *H*2. The figure on the top-left corner reports the type I error generated from $\delta = 0$. The grey horizontal line in the top-left figure represents the 0.05 nominal level. The other figures report the type II errors generated from $\delta \in \{1, 2, 3\}$.

From Figure 8 we make the following observations. First, for type I error, all the three designs have similar performance—all are very close to the 0.05 nominal level. Second, the optimal design almost always has the smallest type II error. This suggests that, even though we design our optimal experiment under the minimax criterion, the optimal design derived from this criterion outperforms the two heuristic benchmarks with respect to the type II error. The type II error becomes smaller when T/m, the effective experimental periods, increases. The gaps between the optimal design and the two heuristic designs also become smaller when T/m increases.

5.5. Estimation Under a Misspecified *m*

5.5.1. Simulation Setup. We run simulations whose setup are similar to Section 5.2.1; the only difference is that we only simulate one assignment path in this Section and conduct hypothesis testing for this single run of the experiment.

The outcome model we consider is in (13); we consider the same parameterization as in Section 5.2.1 and conduct experiments under three different scenarios $\delta \in \{1,2,3\}$. We consider three cases: (i) m correctly specified, so $p = 2$; (ii) $p = 3$, and we estimate the lag-3 causal estimand as in (1); (iii) $p = 1$, and we pretend as if we estimated the lag-1 causal estimand. However, the lag-1 causal estimand is not well defined. Instead, we estimate the 2-misspecified lag-1 causal estimand as in (EC.10) in the online appendix.

We only simulate one assignment path. Because the outcome model is prescribed, we calculate the observed outcomes. There is only one time series of such observed outcomes. We calculate the Horvitz-Thompson estimator based on the simulated assignment path and the simulated observed outcomes. We calculate the lag-p causal estimand directly and also the m-misspecified lag-p causal estimand in conjunction with the simulated assignment path. Finally, we perform the two inference methods from Section 4 and report their associated estimated p-values. For the asymptotic inference method, we plug in $\widehat{\sigma}^2_{\mathrm{U}}$ the conservative upper bound of the variance. We choose $I = 100000$ to be the number of samples drawn in the exact inference method as shown in Algorithm 1.

5.5.2. Simulation Results. Notice this is only one experiment under one simulated experimental setup from one simulated assignment path. So the estimators $\widehat{\tau}_p$ we derive are different from τ_p (or $\tau_p^{(m)}$, which stands for the treatment effect when m is misspecified; see Section EC.4.5 in the online appendix for more details). But they still follow the true causal effects which they estimate. See Table 5.

Table 5. Simulation Results for Correctly Specified m Case and Two Misspecified m Cases

		τ_p	$\tau_p^{(m)}$	$\widehat{\tau}_p$	$\widehat{\sigma}^2_{\mathrm{U}}$	$\widehat{p}_{\mathrm{F}}$	$\widehat{p}_{\mathrm{N}}$
$m=2,p=2$	$\delta=1$	3	–	1.35	8.81	0.626	0.648
	$\delta=2$	6	–	4.30	15.16	0.231	0.269
	$\delta=3$	9	–	7.25	23.88	0.101	0.138
$m=2,p=3$	$\delta=1$	3	–	1.77	14.26	0.606	0.639
	$\delta=2$	6	–	5.00	24.69	0.262	0.314
	$\delta=3$	9	–	8.23	39.00	0.136	0.188
$m=2,p=1$	$\delta=1$	–	2	−1.03	3.87	0.590	0.599
	$\delta=2$	–	4	0.41	6.28	0.866	0.870
	$\delta=3$	–	6	1.86	9.47	0.530	0.547

Notes. The simulation setup for the three $\delta = 1$ cases is the same, so are the $\delta = 2$ cases and $\delta = 3$ cases. The estimated p-values $\widehat{p}_{\mathrm{F}}$ derived from the exact inference are slightly smaller than the p-values $\widehat{p}_{\mathrm{N}}$ derived from the asymptotic inference.

From Table 5, we see that both our estimator and the estimated variance are well defined in all the cases when $p = m$, $p > m$, and $p < m$. In each case, as δ increases from one to three, the associated p-values exhibit decreasing trends, suggesting a stronger rejection rate against the null hypothesis. Moreover, the p-values suggested by the exact inference are always slightly smaller than the p-values suggested by the asymptotic inference. This coincides with our intuition that (i) the exact inference method possesses a stronger null hypothesis (8), which implies the null hypothesis of (9); (ii) in the asymptotic inference, we replaced the true randomization variance by its conservative upper bound, which further leads to a larger p-value.

5.6. Estimation of *m*

We run simulations based on the outcome model as in (13) to test the performance of the procedure described in Section 4.4. In this section, we only focus on $\delta = 3$. Suppose we have narrowed down the range of the order of the carryover effect to be $m \le 3$. In the first round, we use our procedure to test a null hypothesis $m \le 2$. Then we observe rows 3 and 6 from Table 5, with $\widehat{\tau}_2 = 7.25, \widehat{\sigma}^2_2 = 23.88; \widehat{\tau}_3 = 8.23, \widehat{\sigma}^2_3 = 39.00$. So the estimated p-value for the null hypothesis $m \le 2$ is estimated to be $\widehat{p} = 0.902$, which is too large to reject the null hypothesis. In the second round, we consult the procedure to test a null hypothesis $m \le 1$. Then we observe rows 3 and 9 from Table 5, with $\widehat{\tau}_1 = 1.86$, $\widehat{\sigma}^2_3 = 9.47$; $\widehat{\tau}_2 = 7.25, \widehat{\sigma}^2_2 = 23.88$. The estimated p-value for the null hypothesis $m \le 1$ is estimated to be $\widehat{p} = 0.350$. This is still rather large, yet a significant difference from 0.902.

We conduct a few more numerical simulations with different time periods. The setup is the same as in Section 5.5, except that T takes values in $T \in \{210, 1020, 2010\}$.[6] When $T = 210$, in the first round the estimated p-value for the null hypothesis $m \le 2$ is estimated to be $\widehat{p} = 0.956$; in the second round, the estimated p-value for the null hypothesis $m \le 1$ is estimated to be $\widehat{p} = 0.182$. When $T = 1020$, in the first round the estimated p-value for the null hypothesis $m \le 2$ is estimated to be $\widehat{p} = 0.869$; in the second round the estimated p-value for the null hypothesis $m \le 1$ is estimated to be $\widehat{p} = 0.163$. When $T = 2010$, in the first round the estimated p-value for the null hypothesis $m \le 2$ is estimated to be $\widehat{p} = 0.760$; in the second round the estimated p-value for the null hypothesis $m \le 1$ is estimated to be $\widehat{p} = 0.037$. In practice, we suggest increasing the horizon's length to a degree such that $T/p > 100$.

6. Practical Implications, Limitations, and Concluding Remarks

When a firm decides to use a switchback experiment to evaluate a new product or initiative, they have to make multiple decisions to ensure that the results are

reliable, practical, and replicable. First, the firm must determine an appropriate outcome(s) that adequately captures the relative effectiveness of the change. In practice, this requires substantive domain knowledge combined with an understanding of the likely impact of the change; see Kohavi et al. (2020) for an in-depth discussion of metric definition strategies.

Second, as part of the design of the experiment, the firm often has control over the granularity of one single experimental period. As we have shown in Example 7, as long as each time period is smaller than the length of the carryover effect and the length of the carryover effect is divisible by the length of one time unit, the selection of granularity makes no difference to the optimal design and analysis of switchback experiments. On the other hand, setting each period's length longer than the carryover effect will lead to a loss in precision. Consider an extreme case where the carryover effect is one minute, whereas each period is selected to be an hour. If we had set each period to be a minute, we would have collected an order of magnitude of more useful data. Hence, we suggest that each period's length be smaller than the carryover effect duration.

Third, the firm must use prior knowledge to decide an appropriate value p for the order of the carryover effect m. When a firm lacks such knowledge, we propose using the procedure outline in Section 4.4 to select an appropriate value of the order of the carryover effect. Practically, researchers should try to narrow down the set of possible values of m as, when m is relatively large compared with T, our procedure could fail to reject the null hypothesis simply because of insufficient statistical power. Also, it is important to keep in mind that each hypothesis test to identify (12) needs to consume experimental resources at the scale of $T/m > 100$ to distinguish two candidate values, which could be over burdensome when the resource is scarce.

Fourth, when the firm has control over the experiment's horizon, the firm should set $p = m$ and control the overall duration of the experiment $n = T/p = T/m$. We suggest choosing n by referring to the rejection rate curve, as shown in Section 5.3; intuitively, this procedure resembles a typical power analysis. We begin with selecting our inference method, as described in Section 4. We then use our domain knowledge to estimate the expected signal-to-noise ratio; this could be done by looking at historical experiments or through dummy experiments. Then, we choose the desired rejection rate and find out the length of the horizon required.

Finally, using the previous four points, the firm decides the randomization points and samples the assignment path from the appropriate randomization distribution. This final step has already been discussed at length, as we showed in Section 3 the optimal design is obtained from Theorems 1 and 2. In cases when the time horizon is predetermined and when T/p is not an integer, our optimization formulation as shown in Theorem 2 can always be used to find an optimal solution without discarding any periods. Just in the "imperfect cases," we do not have closed-form solutions. Our suggestion is that if the experimental designer wishes not to discard any periods, then solve the optimal solution (using any commercial software); if the experimental designer wishes not to solve an optimization problem, then discard a few periods and consult the explicit solution suggested in Theorem 2.

After designing the experiment, the firm can use the data collected from the test to draw causal conclusions about the new innovation's performance using the two inferential methods as discussed in Section 4. As a more practical consideration, when the firm have the capability to run multiple experiments on multiple experimental units, we suggest the firm to run the optimal design on each of the experimental units and then combine them to increase both precision and power. See Bojinov and Shephard (2019) for detailed discussions.

We point out three limitations of our paper. First, when m, the order of the carryover effect is as large as comparable to T the horizon's length, our method, though still unbiased in theory, incurs a large variance that typically prohibits the firm from making meaningful inference. This is because our method is general and requires the minimum amount of modeling assumptions. If we have strong domain knowledge about the outcome model, we can incorporate it to improve the design. Second, our method only considers flipping independent coins before the experiment even begins. We do not consider adaptively changing the coin flip probabilities, as it requires further assumptions about the outcome model, for example, some time-homogeneity of the data-generating process. Third, in this paper, we have only considered the estimand as in (1), which is motivated when firms want to decide whether to permanently adopt a policy. If the primary focus is on some other general causal estimands, our results do not directly apply. It remains open to derive new results for other estimands, using a similar strategy that we have employed.

Acknowledgments

We thank the department editor George Shanthikumar, the anonymous associate editor, and three anonymous referees whose comments improved the manuscript.

Endnotes

[1] Some authors specifically focus on $p < m$, particularly when m is of the same order as T (Bojinov and Shephard 2019).

[2] When combined with noninterference if there were multiple units, this is known as the stable unit treatment value assumption (Rubin 1980).

[3] Researchers have either shown that versions of completely randomized experiments (corresponding to "fair coin flips") are

optimal, for example, Wu (1981), Li (1983), and Basse et al. (2019b) where they make mild assumptions on permutation invariance, or have explicitly assumed that the coin flips are fair, for example, Bai (2019) and Harshaw et al. (2019).

[4] For other imperfect cases when T is not divisible by m, we can also solve (7) and find the optimal design. However, we do not present closed-form solutions to such a subset selection problem because of integrality issues. Technical discussions about the optimal design in such imperfect cases are deferred to Section EC.3.6 in the online appendix.

[5] We numerically find such variance $\mathsf{Var}(\widehat{\tau}_p)$ and the expectation of the conservative upper bound $\mathbb{E}[\widehat{\sigma}_U^2]$

[6] The values of T were selected such that they were both divisible by both two and three, the possible values of the carryover effect.

References

Abadie A, Athey S, Imbens GW, Wooldridge JM (2020) Sampling-based vs. design-based uncertainty in regression analysis. *Econometrica* 88(1):265–296.

Aronow PM, Samii C (2017) Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Statist.* 11(4):1912–1947.

Athey S, Imbens GW (2018) Design-based analysis in difference-in-differences settings with staggered adoption. Technical report, National Bureau of Economic Research, Cambridge, MA.

Athey S, Eckles D, Imbens GW (2018) Exact p-values for network interference. *J. Amer. Statist. Assoc.* 113(521):230–240.

Azevedo EM, Deng A, Montiel Olea J, Rao JM, Weyl EG (2019) A/b testing with fat tails. *J. Political Econom.* 128(12):4614–4650

Bai Y (2019) Optimality of matched-pair designs in randomized control trials. Preprint, submitted December 15, https://dx.doi.org/10.2139/ssrn.3483834.

Bakshy E, Eckles D, Bernstein MS (2014) Designing and deploying online field experiments. *Proc. 23rd Internat. Conf. World Wide Web* (ACM, New York), 283–292.

Basse G, Ding Y, Toulis P (2019b) Minimax crossover designs. Preprint, submitted August 9, https://arxiv.org/abs/1908.03531v1.

Basse G, Ding P, Feller A, Toulis P (2019a) Randomization tests for peer effects in group formation experiments. Preprint, submitted April 4, https://arxiv.org/abs/1904.02308.

Berger JO (2013) *Statistical Decision Theory and Bayesian Analysis* (Springer Science & Business Media, Berlin).

Bickel PJ, Doksum KA (2015) *Mathematical Statistics: Basic Ideas and Selected Topics*, vol. I (CRC Press, Boca Raton, FL).

Bojinov I, Shephard N (2019) Time series experiments and causal estimands: Exact randomization tests and trading. *J. Amer. Statist. Assoc.* 114(528):1665–1682.

Bojinov I, Rambachan A, Shephard N (2021) Panel experiments and dynamic causal effects: A finite population perspective. *Quant. Econom.* 12(4):1171–1196.

Bojinov I, Saint-Jacques G, Tingley M (2020) Avoid the pitfalls of a/b testing: Make sure your experiments recognize customers' varying needs. *Harvard Bus. Rev.* 98(2):48–53.

Boruvka A, Almirall D, Witkiewitz K, Murphy SA (2018) Assessing time-varying causal effect moderation in mobile health. *J. Amer. Statist. Assoc.* 113(523):1112–1121.

Caro F, Gallien J (2012) Clearance pricing optimization for a fast-fashion retailer. *Oper. Res.* 60(6):1404–1422.

Chamandy N (2016) Experimentation in a ridesharing marketplace—Lyft engineering. Accessed October 1, 2022, https://eng.lyft.com/experimentation-in-a-ridesharing-marketplace-b39db027a66e.

Chamberlain G (1982) Multivariate regression models for panel data. *J. Econometrics* 18(1):5–46.

Chin A (2018) Central limit theorems via Stein's method for randomized experiments under interference. Preprint, submitted April 9, https://arxiv.org/abs/1804.03105.

Cui R, Li J, Zhang D (2020) Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on Airbnb. *Management Sci.* 66(3):1071–1094.

Cui R, Zhang DJ, Bassamboo A (2019) Learning from inventory availability information: Evidence from field experiments on Amazon. *Management Sci.* 65(3):1216–1235.

Deshpande Y, Mackey L, Syrgkanis V, Taddy M (2018) Accurate inference for adaptive linear models. Dy J, Andreas K, eds. *Proc. Internat. Conf. Machine Learn.*, vol. 80 (PMLR), 1194–1203.

Eckles D, Karrer B, Ugander J (2017) Design and analysis of experiments in networks: Reducing bias from interference. *J. Causal Inference* 5(1):20150021.

Farronato C, MacCormack A, Mehta S (2018) Innovation at Uber: The launch of express pool. Harvard Business School Case No. 620(062), Harvard Business School, Boston.

Ferreira KJ, Lee BHA, Simchi-Levi D (2016) Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing Service Oper. Management* 18(1):69–88.

Fisher RA (1937) *The Design of Experiments*, 2nd ed. https://www.amazon.com/Design-Experiments-Ronald-Fisher/dp/0028446909.

Garg N, Nazerzadeh H (2019) Driver surge pricing. Preprint, submitted May 18, https://arxiv.org/abs/1905.07544.

Glynn P, Johari R, Rasouli M (2020) Adaptive experimental design with temporal interference: A maximum likelihood approach. Preprint, submitted June 10, https://arxiv.org/abs/2006.05591.

Gupta S, Kohavi R, Tang D, Xu Y, Andersen R, Bakshy E, Cardin N, et al. (2019) Top challenges from the first practical online controlled experiments summit. *SIGKDD Explorations* 21(1):20–35.

Hadad V, Hirshberg DA, Zhan R, Wager S, Athey S (2019) Confidence intervals for policy evaluation in adaptive experiments. Preprint, submitted November 7, https://arxiv.org/abs/1911.02768v1.

Han KW, Bojinov I, Basse G (2021) Population interference in panel experiments. Preprint, submitted February 28, https://arxiv.org/abs/2103.00553.

Harshaw C, Sävje F, Spielman D, Zhang P (2019) Balancing covariates in randomized experiments using the Gram-Schmidt walk. Preprint, submitted November 8, https://arxiv.org/abs/1911.03071.

Hedayat A, Afsarinejad K (1978) Repeated measurements designs, ii. *Ann. Statist.* 6(3):619–628.

Holtz D, Lobel R, Liskovich I, Aral S (2020) Reducing interference bias in online marketplace pricing experiments. Preprint, submitted April 26, https://arxiv.org/abs/2004.12489.

Imai K, Kim IS (2019) When should we use unit fixed effects regression models for causal inference with longitudinal data? *Amer. J. Political Sci.* 63(2):467–490.

Imbens GW, Rubin DB (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge University Press, Cambridge, UK).

Johari R, Li H, Weintraub G (2020) Experimental design in two-sided platforms: An analysis of bias. Preprint, submitted February 13, https://arxiv.org/abs/2002.05670.

Johari R, Pekelis L, Walsh DJ (2015) Always valid inference: Bringing sequential analysis to a/b testing. Preprint, submitted December 15, https://arxiv.org/abs/1512.04922.

Jones B, Kenward MG (2014) *Design and Analysis of Cross-over Trials* (CRC Press, Boca Raton, FL).

Kastelman D, Ramesh R (2018) Switchback tests and randomized experimentation under network effects at DoorDash. Accessed October 1, 2022, https://medium.com/@DoorDash/switchback-tests-and-randomized-experimentation-under-network-effects-at-doordash-f1d938ab7c2a.

Kempthorne O (1955) The randomization theory of experimental inference. *J. Amer. Statist. Assoc.* 50(271):946–967.

Kohavi R, Thomke S (2017) The surprising power of online experiments. *Harvard Bus. Rev.* 95:74–82.

Kohavi R, Henne RM, Sommerfield D (2007) Practical guide to controlled experiments on the web: Listen to your customers not to the hippo. *Proc. 13th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 959–967.

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

Kohavi R, Crook T, Longbotham R, Frasca B, Henne R, Ferres JL, Melamed T (2009) Online experimentation at Microsoft. *Data Mining Case Stud.* 11:39.

Koning R, Hasan S, Chatterji A (2019) Experimentation and startup performance: Evidence from a/b testing. Technical report, National Bureau of Economic Research, Cambridge, MA.

Laird NM, Skinner J, Kenward M (1992) An analysis of two-period crossover designs with carry-over effects. *Statist. Medicine* 11(14-15):1967–1979.

Li H, Zhao G, Johari R, Weintraub GY (2021) Interference, bias, and variance in two-sided marketplace experimentation: Guidance for platforms. Preprint, submitted April 25, https://arxiv.org/abs/2104.12222.

Li JQ, Rusmevichientong P, Simester D, Tsitsiklis JN, Zoumpoulis SI (2015) The value of field experiments. *Management Sci.* 61(7):1722–1740.

Li KC (1983) Minimaxity for randomized designs: Some general results. *Ann. Statist.* 11(1):225–239.

Li X, Ding P (2017) General forms of finite population central limit theorems with applications to causal inference. *J. Amer. Statist. Assoc.* 112(520):1759–1769.

Li X, Ding P, Rubin DB (2020) Rerandomization in 2^k factorial experiments. *Ann. Statist.* 48(1):43–63.

Lillie EO, Patay B, Diamant J, Issell B, Topol EJ, Schork NJ (2011) The n-of-1 clinical trial: The ultimate strategy for individualizing medicine? *Personalized Medicine* 8(2):161–173.

Ma W, Simchi-Levi D, Zhao J (2021) Dynamic pricing (and assortment) under a static calendar. *Management Sci.* 67(4):2292–2313.

March JG (1991) Exploration and exploitation in organizational learning. *Organ. Sci.* 2(1):71–87.

McFowland III E, Somanchi S, Neill DB (2018) Efficient discovery of heterogeneous treatment effects in randomized experiments via anomalous pattern detection. Preprint, submitted March 24, https://arxiv.org/abs/1803.09159.

Neyman J, Dabrowska DM, Speed TP (1990) On the application of probability theory to agricultural experiments: Essay on principles, section 9. *Statist. Sci.* 5(4):465–472.

Nie X, Tian X, Taylor J, Zou J (2018) Why adaptively collected data have negative bias and how to correct for it. Storkey A, Perez-Cruz F, eds. *Proc. 21st Internat. Conf. Artificial Intelligence Statist.*, vol. 84 (PMLR), 1261–1269.

Oman SD, Seiden E (1988) Switch-back designs. *Biometrika* 75(1):81–89.

Puelz D, Basse G, Feller A, Toulis P (2019) A graph-theoretic approach to randomization tests of causal effects under general interference. Preprint, submitted October 24, https://arxiv.org/abs/1910.10862.

Rambachan A, Shephard N (2019) Econometric analysis of potential outcomes time series: Instruments, shocks, linearity and the causal response function. Preprint, submitted March 5, https://arxiv.org/abs/1903.01637.

Robins J (1986) A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Math. Model.* 7(9-12):1393–1512.

Rubin DB (1980) Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* 75(371):591–593.

Sarasvathy SD (2001) Causation and effectuation: Toward a theoretical shift from economic inevitability to entrepreneurial contingency. *Acad. Management Rev.* 26(2):243–263.

Senn S, Lambrou D (1998) Robust and realistic approaches to carry-over. *Statist. Medicine* 17(24):2849–2864.

Sitkin SB (1992) Learning through failure: The strategy of small losses. *Res. Organ. Behav.* 14:231–266.

Sobel ME (2012) Does marriage boost men's wages?: Identification of treatment effects in fixed effects regression models for panel data. *J. Amer. Statist. Assoc.* 107(498):521–529.

Sun T, Viswanathan S, Huang N, Zheleva E (2018) Designing promotional incentive to embrace social sharing: Evidence from field and laboratory experiments. Preprint, submitted January 5, https://dx.doi.org/10.2139/ssrn.3095094.

Sussman DL, Airoldi EM (2017) Elements of estimation theory for causal effects in the presence of network interference. Preprint, submitted February 12, https://arxiv.org/abs/1702.03578.

Thomke S (2001) Enlightened experimentation: The new imperative for innovation. *Harvard Bus. Rev.* 79(2):66–75.

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

Wager S, Xu K (2019) Experimenting in equilibrium. Preprint, submitted March 6, https://arxiv.org/abs/1903.02124.

Wu CF (1981) On the robustness and efficiency of some randomized designs. *Ann. Statist.* 9(6):1168–1177.

Xiong R, Athey S, Bayati M, Imbens GW (2019) Optimal experimental design for staggered rollouts. Preprint, submitted November 9, https://arxiv.org/abs/1911.03764v1.