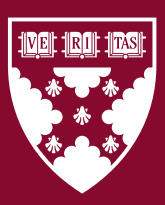
Working Paper 24-075

# Winner Take All: Exploiting Asymmetry in Factorial Designs

Matthew DosSantos DiSorbo
Iavor Bojinov
Fiammetta Menchetti



Harvard Business School

# Winner Take All: Exploiting Asymmetry in Factorial Designs

Matthew DosSantos DiSorbo
Harvard Business School

Iavor Bojinov
Harvard Business School

Fiammetta Menchetti
University of Florence

**Working Paper 24-075**



Funding for this research was provided in part by Harvard Business School.

# Winner Take All: Exploiting Asymmetry in Factorial Designs

Matthew DosSantos DiSorbo[1] Iavor Bojinov[1] Fiammetta Menchetti[2]

[1]Harvard Business School
{mdisorbo, ibojinov}@hbs.edu
[2]University of Florence
fiammetta.menchetti@unifi.it

**Abstract**

Researchers and practitioners have embraced factorial experiments to simultaneously test multiple treatments, each with different levels. With the rise of technologies like Generative AI, factorial experimentation has become even more accessible: it is easier than ever to generate different versions of potential treatments. Typically, in large-scale factorial experiments, the primary objective is to identify the treatment with the largest causal effect. This is especially true for experiments that suffer from measurement error, attrition, non-compliance, and censoring: point estimates are unreliable, but — as we show — the asymmetry in the largest treatment effect makes it possible to identify the most impactful treatment even when point estimates are biased. To exploit this asymmetry, we propose using a Fisher randomization test as a general non-parametric approach for inference, which we apply to an existing field experiment that measured intern performance at a large financial firm. We show that the earliest possible intervention has an immediate and enduring impact: performance improves in the week of the intervention and in future weeks, sometimes even to a greater extent than interventions in those future weeks. The takeaway — intervene early — has important consequences across the many contexts of workplace programs.

**Keywords:** factorial designs, Fisher randomization, rank estimators, employer interventions, causal inference

## 1 Introduction

Within academia and, more recently, industry, experimentation has become the gold standard for determining the causal effects of treatments on outcomes of interests (Levitt and List 2009, Kohavi et al. 2013, 2020, Thomke 2020). Technology companies in particular have embraced experimentation as part of the innovation process: experiments are now regularly used to determine the effect of a website's color scheme on consumer buying intention (Pelet and Papadopoulou 2012), measure

how personification impacts advertising performance (Salminen et al. 2021), and identify ways to improve human-algorithm collaboration (Sun et al. 2022). Studies indicate that implementing experimentation technology and adopting rigorous scientific frameworks help startups perform better (Camuffo et al. 2020, Koning et al. 2022), while iterative experimentation at larger companies can also improve key metrics (Mao and Bojinov 2021). In academia, experimentation has a long history in the hard sciences and has recently been adopted by empirical researchers in management science; for example, researchers have used experiments to estimate the impact of operational transparency (Buell et al. 2017) and determine methods to generate the best idea (Girotra et al. 2010).

In complex experiments, many treatments can be tested simultaneously; modern tools like generative AI make these designs more accessible than ever before. Consider a marketing experiment that tests the impact of important factors like heading, image, and color palette. A stable diffusion model (Zhang et al. 2023) can quickly generate a large number of advertisements, each with multiple factors and levels[1], which are shown to potential customers in different treatment groups. In such an experiment, the click-through rate — the number of times an ad is clicked, divided by the number of times it is shown — is often a primary outcome that is estimated for different levels of each treatment. However, researchers might also be interested in conditional conversion rate, or the number of sales divided by the number of participants who actually visit the site. In this experiment, sales outcomes are defined conditionally and thus are not always observed: individuals who do not click the ad do not visit the site, and the researcher does not know if they would, or would not, have made a purchase. Estimates of the advertisement's true effect on conditional conversion rate will be biased if the population that clicks the ad has different purchasing behavior than the population that does not click the ad. Unfortunately, there are many forces in addition to conditional outcomes that can corrupt estimates in randomized experiments: attrition (Hewitt et al. 2010), measurement error (Gillen et al. 2019), treatment noncompliance (Sagarin et al. 2014) and data censoring (Rubin 2006), to name a few.

Researchers have proposed methods to overcome these challenges. Most rely on extensive covariate data, and others on intensive follow-up with experimental subjects. One can use baseline covariates to estimate compliance behavior (Jin and Rubin 2008), conduct intensive qualitative work to measure the impact of misreporting on sensitive topics (Blattman et al. 2016) or rerandomize to achieve improved balance across covariates (Branson et al. 2016). Unfortunately, these strategies can be infeasible in practical settings. For example, the experiment discussed in this paper centers on interns at a large financial firm: we have sparse covariate data, no access to the interns for follow-up discussions, and no ability to rerandomize treatment assignment. More generally, experiments with corporate partners may lack extensive covariates *and* follow-up opportunities, as firms may be concerned about how full data access impacts trust (Radulescu 2018).

To address these challenges, we employ a simple idea that leverages the asymmetry of the most impactful treatment: if the point estimate of the largest effect is biased upwards, it is even easier to

[1]For example, a treatment could be image size with levels small, medium and large.

identify that treatment. We focus on settings with factorial experiments, which trace their origins to 20th century agriculture (Yates 1937, Fisher 1942), and have recently surged in popularity across fields (Dasgupta et al. 2015, Dong 2015, Egami and Imai 2018, Espinosa et al. 2016, Lu 2016, Lu and Deng 2017, Mukerjee et al. 2018, Zhao et al. 2018, Pashley and Bind 2023). Factorial designs allow researchers to assess multiple treatment effects; for instance, measuring the impact of five different incentive programs on school performance (Branson et al. 2016, Dasgupta et al. 2015). Unfortunately, point estimates can be untrustworthy: we define and derive the bias in factorial design estimates under measurement error and censored outcomes, and show that this bias is non-negligible in real world settings. However, while these biases can disrupt point estimates, we illustrate how our method exploits the asymmetry of the largest effect to reliably identify the most impactful treatment, and introduce flexible Fisher randomization tests to perform this identification. Put simply, in this paradigm, not all bias is harmful: positive bias can help us determine the largest treatment effect.

Identifying the most impactful feature has great practical importance. Firms often practice "Opportunity Sizing", analyzing a set of treatments to identify possible candidates for implementation (Bojinov et al. 2020). Although this is common in the causal inference literature, the same principle applies for experimentation: interventions are costly, and experiments can help practitioners determine which of many 'levers' will have the greatest impact. In turn, it may be sufficient to identify the most impactful treatment without recovering the exact treatment effects. We implement our method by deriving insights from a randomized experiment conducted by researchers in collaboration with a major financial institution (Bojinov et al. 2021). The partner firm ran a five-week virtual internship program in the summer of 2020 for $1,370$ remote interns. Interns were assigned to various treatments, which included a virtual intern-senior manager water cooler (videoconference) between a group of interns and a senior manager who did not regularly interact with interns. Interns were assigned to attend the water cooler chats in some combination of weeks across the time horizon of the experiment. The researchers concluded that these virtual water coolers may deliver improved outcomes for interns when they create a demographic match between interns and senior managers, or occur at regular intervals.

We extend these insights by considering the *timing* of treatment, i.e., identifying the week in which the videoconference had the largest impact on intern performance. This leverages a factorial design, where multiple treatments are applied in different combinations across subjects (Fisher 1942, Yates 1937, Dasgupta et al. 2015). We consider the different weeks of the intervention as distinct treatments because we believe that applying treatment at various points in the intern cycle will yield different effects on intern performance. The critical outcome — performance rating, assigned by the intern's manager during each week of the internship — is subject to severe rounding error, since managers can only assign ratings on a coarse scale (1, 2 or 3). While estimates of how much each treatment impacts an intern's performance will be biased by this imprecision, it may still be possible to identify the week where the intervention is most impactful. We conclude that the earliest possible

interventions have an immediate and enduring positive impact on intern outcomes; for instance, treatment in the first week has a large impact on future weeks, larger even than treatment applied in those future weeks.

Another critical outcome is offer acceptance: the likelihood that an intern accepts an extended offer. Unfortunately, this data is censored: we do not observe the outcome (accept an offer, or not) for interns that do not *receive* an offer. Once again, point estimates will likely be biased — we derive the bias for factorial design estimates under conditional outcomes which is, to our knowledge, a novel contribution — but it may be possible to identify the most impactful treatment. Our results were, once again, suggestive that early interventions increase the likelihood of accepting an extended offer. Ultimately, these results were not statistically significant: this is a low variance setting, since most interns accepted the extended offer. However, we hope that our approach can be leveraged across the many managerial settings where attrition, data censoring and noncompliance hinder estimation.

Our research contributes to the field of design-based methods, including terminating experiments without a fixed time horizon (Ham et al. 2022), addressing problems of interference (Bojinov and Shephard 2019) and rerandomization (Li et al. 2020), proposing designs for switchback (Bojinov et al. 2022) and split-plot (Zhao and Ding 2022) experiments, accounting for design-based uncertainty in standard errors (Abadie et al. 2020) and addressing unexplained treatment effect variation (Ding et al. 2016). We also add to the literature that measures the efficacy of different workplace programs and initiatives (Hart et al. 2019, Ellenbecker et al. 2007, Vicente-Herrero et al. 2004, Galván Vela et al. 2022, Kim et al. 2016, Cluff et al. 2018, Spence 2015, Mujtaba and Cavico 2013). However, to our knowledge, this is the first study that examines the proper *timing* of a workplace intervention. We believe these results deliver fresh insights to workplace initiatives as common and well-studied as corporate wellness programs (Gubler et al. 2018, Mills et al. 2007, Ni Mhurchu et al. 2010) and flexible working arrangements (Maxwell et al. 2007).

The paper is organized as follows. Section 2 reviews factorial designs, while Section 3 covers estimation and testing. Section 4 details common sources of experimental bias and their impact on factorial designs. Section 5 conducts a simulation study designed to test our ability to identify the most impactful feature. Finally, Section 6 extracts insights from a large managerial experiment.

# 2 Factorial Design

In the experiment we analyze, the intervention — a water-cooler chat with a senior manager — can be applied in any combination of weeks across the internship. We define the 'treatment' as receiving the intervention in a specific week; therefore, a single intern may receive multiple treatments (i.e., water-cooler chats in Weeks 2 and 4) and these multiple treatments constitute a single treatment allocation. In Section 2.1, we revise the notation associated with factorial designs; in Section 2.2, we introduce ranking estimands to the factorial paradigm. Section 2.3 extends to fractional factorial designs, where only a subset of treatment combinations is represented in the experiment.

## 2.1 Treatments and potential outcomes

In a $2^K$ full factorial design, $N$ units are exposed to $K$ treatments with two levels each, thereby defining $2^K = J$ possible treatment combinations for each unit. In our application, the units are summer interns, and the treatments are different weeks that the intervention (a water cooler chat with a senior manager) may be applied. A 'treatment combination' is a specific allocation of different treatment weeks. For example, one treatment combination might consist of interns who received an intervention in Weeks 2, 4, and 5 of the internship.

Following Dasgupta et al. (2015), we denote by $\boldsymbol{z}_j = (z_{j,1}, \ldots, z_{j,K})$ the $j$-the treatment combination, with $z_{j,k} \in \{-1, 1\}$ being the level of factor $k$ in the $j$-th combination. Table 1 gives an example with three weeks; for instance, $\boldsymbol{z}_4 = (1, -1, -1)$ indicates that the virtual water cooler chat with a senior manager was assigned only during the first week of the internship. We also indicate with $W_i(\boldsymbol{z}_j) \in \{0, 1\}$ the treatment assignment of intern $i$, where 1 indicates that the intern receives the treatment combination $\boldsymbol{z}_j$ and 0 indicates that the intern receives another treatment combination.

Throughout, we assume that each intern has $J$ potential outcomes that they would exhibit corresponding to each treatment combination (Splawa-Neyman et al. 1990, Rubin 1974). That is, $Y_i(\boldsymbol{z}_j)$ indicates the outcome (i.e., performance rating) for intern $i$ under treatment allocation $j$. Then $\bar{Y}(\boldsymbol{z}_j)$ gives the average outcome under treatment allocation $j$ across interns, and $\bar{\boldsymbol{Y}}(\boldsymbol{z}) = (\bar{Y}(\boldsymbol{z}_1), ..., \bar{Y}(\boldsymbol{z}_{2^K}))$. Again, we only observe one of the potential outcomes for each intern: the outcome (performance rating) under the treatment allocation (weeks of treatment) they received.

Implicitly, we are assuming that there is no interference between experimental subjects; that is, the treatment of one unit cannot affect the outcome of another (Cox 1958). Often, researchers combine the no-inference assumption with a restriction that every unit receives the same version of the treatment. Together, these restrictions are commonly known as the Stable Unit Treatment Value Assumption (Rubin 1980) or SUTVA. In our case, it is unlikely that the treatment for one intern would impact another intern's performance, as the internship was conducted remotely. In addition, we can safely exclude the possibility that there were different versions of the treatment leading to different potential outcomes. Therefore, we believe that SUTVA is reliable in our empirical application.

## 2.2 Causal Estimands

The main effect of feature $k$ can be defined by contrasting the potential outcomes when treatment $k$ is "on" (included in a treatment combination) to those when it is "off" (not included in a treatment combination). For ease of exposition, we use the $J$ lengthed vector $\boldsymbol{g}_k$, whose $k^{th}$ entries take a value of $+1$ if the $k^{th}$ treatment is on and $-1$ otherwise. In our example, $\boldsymbol{g}_k$ indicates the weeks in which treatment was received or was not received. From Dasgupta et al. (2015), the main effect of treatment $k$ is then

$$\tau(k) = \frac{1}{2^{K-1}} \boldsymbol{g}_k' \bar{\boldsymbol{Y}}(\boldsymbol{z}). \tag{1}$$

Let $\boldsymbol{\tau} = (\tau(1), \ldots, \tau(K))$ be the vector containing the main effects of the $K$ factors and denote with $\tau_1, \ldots, \tau_K$ the corresponding order statistics, such that $\tau_1 = \min\{\tau(1), \ldots, \tau(K)\}$ and $\tau_K = \max\{\tau(1), \ldots, \tau(K)\}$. We say that $\tau(k)$ has rank $R_k = h$ if $\tau(k) = \tau_h$, with $h = 1, \ldots, K$. Thus, the **rank of the main effect vector** is defined as

$$R(\boldsymbol{\tau}) = (R_1, \ldots, R_K) \text{ s.t } \tau(k) = \tau_h \;\; h, k = 1, \ldots, K. \tag{2}$$

In our context, $R_k = K$ implies that an intervention in week $k$ has the largest effect on intern outcomes; $R_k = 1$ implies that an intervention in week $k$ has the smallest effect. Firms may be more interested in the former: in our example, the treatment can be applied at many times during the internship, and it is important to determine which week yields the largest effect. This is more practically relevant than determining the week with the smallest or second-largest impact. Since interventions are costly and require time commitment from senior managers, the company's goal was to identify a single week in which the intervention had a maximum effect. Fortunately, as we will discuss in later sections, it is easier to identify the feature with rank $K$ because we cannot overestimate the rank of the most impactful feature; we can only underestimate it.

Full Factorial Design.

| Treatment | Week 1 | Week 2 | Week 3 | Potential Outcomes | Assignment | Average |
|---|---|---|---|---|---|---|
| $\boldsymbol{z}_1$ | 1 | 1 | 1 | $Y_i(\boldsymbol{z}_1)$ | $W_i(\boldsymbol{z}_1)$ | $\bar{Y}(\boldsymbol{z}_1)$ |
| $\boldsymbol{z}_2$ | 1 | 1 | -1 | $Y_i(\boldsymbol{z}_2)$ | $W_i(\boldsymbol{z}_2)$ | $\bar{Y}(\boldsymbol{z}_2)$ |
| $\boldsymbol{z}_3$ | 1 | -1 | 1 | $Y_i(\boldsymbol{z}_3)$ | $W_i(\boldsymbol{z}_3)$ | $\bar{Y}(\boldsymbol{z}_3)$ |
| $\boldsymbol{z}_4$ | 1 | -1 | -1 | $Y_i(\boldsymbol{z}_4)$ | $W_i(\boldsymbol{z}_4)$ | $\bar{Y}(\boldsymbol{z}_4)$ |
| $\boldsymbol{z}_5$ | -1 | 1 | 1 | $Y_i(\boldsymbol{z}_5)$ | $W_i(\boldsymbol{z}_5)$ | $\bar{Y}(\boldsymbol{z}_5)$ |
| $\boldsymbol{z}_6$ | -1 | 1 | -1 | $Y_i(\boldsymbol{z}_6)$ | $W_i(\boldsymbol{z}_6)$ | $\bar{Y}(\boldsymbol{z}_6)$ |
| $\boldsymbol{z}_7$ | -1 | -1 | 1 | $Y_i(\boldsymbol{z}_7)$ | $W_i(\boldsymbol{z}_7)$ | $\bar{Y}(\boldsymbol{z}_7)$ |
| $\boldsymbol{z}_8$ | -1 | -1 | -1 | $Y_i(\boldsymbol{z}_8)$ | $W_i(\boldsymbol{z}_8)$ | $\bar{Y}(\boldsymbol{z}_8)$ |
| $\boldsymbol{z}$ | $\boldsymbol{g}_1$ | $\boldsymbol{g}_2$ | $\boldsymbol{g}_3$ | $Y_i(\boldsymbol{z})$ | $W_i(\boldsymbol{z})$ | $\bar{\boldsymbol{Y}}(\boldsymbol{z})$ |

## 2.3 Fractional factorial experiments

Full factorial designs require that units are assigned to all possible treatment allocations. Therefore, in practice, as the number of treatment combinations grows exponentially, running a full factorial experiment can become costly and infeasible. For example, weekly treatments over a 10-week internship would require $2^{10} = 1,024$ different treatment combinations. If there are an average of 30 interns per treatment combination, then more than $30,000$ interns would be required to conduct the experiment.

In these cases, fractional factorial designs may be preferred. A $2^{K-p}$ fractional factorial design is constructed by writing the full factorial for $K - p$ factors and substituting the remaining $p$ columns with a multiplicative combination of the existing factor columns. Table 2 shows a fractional

factorial design $2^{3-1}$, which represents a one-half fraction of the full $2^3$ design considered above. The contrast columns $\boldsymbol{g}_k^*$ are shortened versions of $\boldsymbol{g}_k$ in a full factorial design; similarly, the treatment combinations $\boldsymbol{z}^*$ are a subset of all the combinations in the $2^3$ design. A more general concept is an incomplete factorial design, where some treatment combinations are removed, perhaps because they are infeasible to apply simultaneously (Byar et al. 1993).

Unfortunately, while fractional factorial designs require smaller sample sizes, they do not precisely extract the main effect of each feature. To understand this, consider the fractional factorial design in Table 2. As discussed in Montgomery (2017), this design is said to have the defining relation $I = ABC$ where $I$ is the identity and $A, B$ and $C$ represent the $\pm 1$ columns for interventions in Weeks 1, 2 and 3, respectively. If we multiply the defining relation by $A$, we obtain $A = A^2BC = BC$, since $A^2 = I$. This implies that $A = BC$, or the main effect of an intervention in Week 1 is indistinguishable from the interaction $BC$. This $BC$ interaction is defined as the average effect of an intervention in Week 2 when there is also an intervention in Week 3, minus the average effect of an intervention in Week 2 when there is no intervention in Week 3. Ultimately, this means that the effect of $A$ — an intervention in Week 1 — cannot be distinguished from that of $BC$, the interaction of interventions in Weeks 2 and 3. When we are estimating the effect of treatment $A$ in a $2^{3-1}$ fractional factorial design, what we really estimate is $A + BC$. When multiple effects share this property, they are considered *aliases* (Montgomery 2017).

Fractional Factorial Design.

| Treatment | Week 1 (A) | Week 2 (B) | *Week 3 (C)* | Potential Outcomes | Assignment | Average |
|---|---|---|---|---|---|---|
| $\boldsymbol{z}_1^*$ | 1 | 1 | 1 | $Y_i(\boldsymbol{z}_1^*)$ | $W_i(\boldsymbol{z}_1^*)$ | $\bar{Y}(\boldsymbol{z}_1^*)$ |
| $\boldsymbol{z}_2^*$ | 1 | -1 | -1 | $Y_i(\boldsymbol{z}_2^*)$ | $W_i(\boldsymbol{z}_2^*)$ | $\bar{Y}(\boldsymbol{z}_2^*)$ |
| $\boldsymbol{z}_3^*$ | -1 | 1 | -1 | $Y_i(\boldsymbol{z}_3^*)$ | $W_i(\boldsymbol{z}_3^*)$ | $\bar{Y}(\boldsymbol{z}_3^*)$ |
| $\boldsymbol{z}_4^*$ | -1 | -1 | 1 | $Y_i(\boldsymbol{z}_4^*)$ | $W_i(\boldsymbol{z}_4^*)$ | $\bar{Y}(\boldsymbol{z}_4^*)$ |
| $\boldsymbol{z}^*$ | $\boldsymbol{g}_1^*$ | $\boldsymbol{g}_2^*$ | $\boldsymbol{g}_3^*$ | $Y_i(\boldsymbol{z}^*)$ | $W_i(\boldsymbol{z}^*)$ | $\bar{\boldsymbol{Y}}(\boldsymbol{z}^*)$ |

# 3 Estimation and testing

The quantities defined in Section 2 can be estimated using popular, well-defined estimators (Dasgupta et al. 2015, Pashley and Bind 2023). These allow us to extract estimates of individual treatment effects; in our setting, the impact of intervening during each week of the internship. We go further by introducing, in Section 3.1, estimators for the ranks of the individual effects, which allows to answer which week has the most impactful intervention. Then, we introduce methods to test the estimators in Section 3.2 and argue how this approach can mitigate the bias introduced to traditional point estimates in fractional factorial designs in Section 3.3.

## 3.1 Rank estimators

Let $\bar{\boldsymbol{Y}}^{obs}(\boldsymbol{z}) = (\bar{Y}^{obs}(\boldsymbol{z}_1), \ldots, \bar{Y}^{obs}(\boldsymbol{z}_J))$ be the vector of observed mean potential outcomes for each treatment. From Dasgupta et al. (2015), an estimator of the effect in Equation 1 is given by

$$\hat{\tau}(k) = \frac{1}{2^{K-1}} \boldsymbol{g}_k' \bar{\boldsymbol{Y}}^{obs}(\boldsymbol{z}). \tag{3}$$

In our empirical example, these estimates allow us to measure the impact of treatments on intern outcomes. Again, we can introduce estimators for the effect ranks.

Let $\hat{\boldsymbol{\tau}} = (\hat{\tau}(1), \ldots, \hat{\tau}(K))$ be the vector of the main effect estimators. An estimator for the rank of $\hat{\tau}_k$ can be defined as the number of times that $\hat{\tau}(k)$ exceeds the other estimated effects in $\hat{\boldsymbol{\tau}}$, namely

$$\hat{R}_k = \sum_{h=1}^{K} \mathbb{1}_{\hat{\tau}(k) \geq \hat{\tau}(h)}. \tag{4}$$

Then, the ranking of the main effect estimators is the K-dimensional vector such that the $k$-th element corresponds to $\hat{R}_k$:

$$R(\hat{\boldsymbol{\tau}}) = (\hat{R}_1, \ldots, \hat{R}_K). \tag{5}$$

## 3.2 Fisher Testing

We now introduce a Fisher randomization test for a sharp null hypothesis that can be easily implemented in a variety of settings.

**Hypothesis 1**

$$H_0 : \; Y_i(\boldsymbol{z}_1) = Y_i(\boldsymbol{z}_2) = ... = Y_i(\boldsymbol{z}_J).$$

Note that this null hypothesis implies, using Equation 1, that $\tau(1) = ... = \tau(K)$, or that there is no difference between the main effects in the full factorial design.

Recall that only one of the potential outcomes is observed. Under Hypothesis 1, all potential outcomes across $K$ treatment combinations are equal to the observed outcome. These dynamics are depicted for a $2^2$ full factorial design in Table 3: the first group of columns illustrates how only one potential outcome is observed, and the second group of columns depicts how the sharp null assumes that all potential outcomes are equal. To conduct the test, we can assume the null hypothesis is correct and sample alternative assignment matrices; one example resampling is shown in the final group of columns. For each resampling, the estimates of the main effects can be recovered from Equation 3.

Let $\hat{\tau}_K$ be the estimate of the effect with estimated rank $K$; that is, the largest point estimate. Then we can write the test statistic — for testing the largest point estimate — as

$$T = \hat{\tau}_K - \hat{\tau}_{K-1}. \tag{6}$$

| | Potential Outcomes | | | | Sharp Null Hypothesis | | | |
|---|---|---|---|---|---|---|---|---|
| Unit | $Y_i(\boldsymbol{z}_1)$ | $Y_i(\boldsymbol{z}_2)$ | $Y_i(\boldsymbol{z}_3)$ | $Y_i(\boldsymbol{z}_4)$ | $Y_i(\boldsymbol{z}_1)$ | $Y_i(\boldsymbol{z}_2)$ | $Y_i(\boldsymbol{z}_3)$ | $Y_i(\boldsymbol{z}_4)$ |
| 1 | ? | 2.4 | ? | ? | 2.4 | 2.4 | 2.4 | 2.4 |
| 2 | 7.8 | ? | ? | ? | 7.8 | 7.8 | 7.8 | 7.8 |
| 3 | ? | ? | ? | 9.2 | 9.2 | 9.2 | 9.2 | 9.2 |
| 4 | ? | ? | 3.2 | ? | 3.2 | 3.2 | 3.2 | 3.2 |
| 5 | 1.2 | ? | ? | ? | 1.2 | 1.2 | 1.2 | 1.2 |
| 6 | ? | ? | ? | 8.1 | 8.1 | 8.1 | 8.1 | 8.1 |

| | Example Resampling | | | |
|---|---|---|---|---|
| Unit | $Y_i(\boldsymbol{z}_1)$ | $Y_i(\boldsymbol{z}_2)$ | $Y_i(\boldsymbol{z}_3)$ | $Y_i(\boldsymbol{z}_4)$ |
| 1 | ? | ? | 2.4 | ? |
| 2 | ? | 7.8 | ? | ? |
| 3 | 9.2 | ? | ? | ? |
| 4 | ? | 3.2 | ? | ? |
| 5 | ? | ? | ? | 1.2 |
| 6 | ? | ? | ? | 8.1 |

Table 3: Potential outcomes in a $K = 2$ full factorial design. For each unit, only one potential outcome is observed. For example, $W_6(\boldsymbol{z}_4) = 1$, since unit 6 receives treatment 4. In our sharp null hypothesis, every potential outcome is equal to the observed outcome for every unit. In each resampling, units are assigned to a different treatment, under the sharp null hypothesis that all potential outcomes are equal.

The p-value of our test is $P(T^{obs} > T)$, where $T^{obs}$ is the observed difference between $\hat{\tau}_K$ and $\hat{\tau}_{K-1}$. If the test is significant, we can extend the paradigm to find a Fisher confidence interval. The new test statistic becomes:

$$T_{CI} = \hat{\tau}_K - \hat{\tau}_{K-1} - c.$$

The significant values of $c$ constitute the confidence interval for the difference between $\hat{\tau}_K$ and $\hat{\tau}_{K-1}$. For example, an interval of $[-\infty, 0.07)$ with $\alpha = 0.10$ implies that we have 90% confidence that the $\hat{\tau}_K$ is at least 0.07 larger than $\hat{\tau}_{K-1}$.

Finally, Hypothesis 1 also implies that $\tau(1) = \tau(2) = ... = \tau(K) = 0$, since if there is no difference between potential outcomes then none of the treatment allocations have an effect. We can thus leverage another test statistic

$$T_0 = \hat{\tau}_i,$$

which tests if main effect $i$ is non-zero. The p-value can be calculated as $P(T_0^{obs} > T_0)$.

This approach is purposefully general: it can apply in experiments impacted by measurement error, attrition, data censoring and noncompliance. Recent work by Aronow et al. (2024) discusses another test statistic for randomization-based confidence intervals — a "studentized Anderson—Rubin-type statistic" — specifically for experiments facing noncompliance.

### 3.3 Ranks in Fractional designs

The definition of the main effect estimators $\hat{\tau}$, as well as their ranks $\hat{R}$, can be easily generalized from full factorial designs to a fractional design setting. An estimator for (1) is given by

$$\hat{\tau}^*(k) = \frac{1}{2^{K-p-1}} \boldsymbol{g}_k^{*'} \bar{\boldsymbol{Y}}^{obs}(\boldsymbol{z}^*). \tag{7}$$

Further, by ranking the estimated effect we can obtain an estimate of the true ranking (2). Let $\hat{\boldsymbol{\tau}}^* = (\hat{\tau}^*(1), \ldots, \hat{\tau}^*(K))$ be the vector of the main effect estimator in a fractional factorial design and indicate with $\hat{R}_k^*$ the rank of $\hat{\tau}^*(k)$ for each $k = 1, \ldots, K$. An estimator for the rank of $\hat{\tau}^*(k)$ can be defined akin to an order statistic, as the number of times that $\hat{\tau}^*(k)$ exceeds the estimated effects in $\hat{\boldsymbol{\tau}}^*$. Namely,

$$\hat{R}_k^* = \sum_{h=1}^{K} \mathbb{1}_{\hat{\tau}^*(k) \geq \hat{\tau}^*(h)}. \tag{8}$$

The ranking of the main effect estimators is the K-dimensional vector such that the $k$-th element corresponds to $\hat{R}_k^*$,

$$R(\hat{\boldsymbol{\tau}}^*) = (\hat{R}_1^*, \ldots, \hat{R}_K^*). \tag{9}$$

Note that the estimates in fractional designs are not necessarily unbiased. Following Pashley and Bind (2023), we can write the bias of the estimator in Equation 7:

$$B(k) = \sum_{j \in \mathcal{A}[k] \setminus \{k\}} (-1)^{A_{k,j}} \tau(j), \tag{10}$$

where $\mathcal{A}[k]$ is the set of features and interactions that are aliased with feature $k$ and $A_{k,j}$ is an indicator for feature $j$ being negatively aliased with feature $k$. In other words, in a fractional factorial design, the main effects are biased by the size of the interactions aliasing with them.

A common assumption is to assume that the $\tau(j)$ effects are negligible, especially if they are composed of higher-order interactions (Montgomery 2017, Wu and Hamada 2000). For example, in a $2^{5-1}$ fractional factorial design with defining relation $I = ABCDE$, the first treatment is aliased by the four-treatment interaction $BCDE$. Recall that the $BC$ interaction is defined as the difference between the average effect of $B$ when $C$ is on and the average effect of $B$ when $C$ is off. Going further, $BCD$ is the average difference between the $BC$ interaction when $D$ is on or off, and $BCDE$ is the average difference between the $BCD$ interaction when $E$ is on or off.

Clearly, with each additional treatment, the compound interaction becomes more granular, which informs the assumption that larger interactions are negligible. However, this is an active assumption that the experimenter must make and may not always hold. Conversely, the Fisher test statistic in Equation 6 may retain its important properties even when aliasing interactions introduce bias in point estimates.

**Remark 1** *WLOG, let $\tau(k)$ be the largest and $\tau(h)$ be the second-largest treatment effects. Since $\hat{\tau}(k)$ is biased by $B(k)$ for $\tau(k)$, we can write*

$$E[\hat{\tau}(k) - \hat{\tau}(h)] = \tau(k) - \tau(h) + E[B(k) - B(h)].$$

*Note that $\tau(k) - \tau(h) > 0$, and if $\tau(k) - \tau(h) > E[B(h) - B(k)]$, then $E[\hat{\tau}(k) - \hat{\tau}(h)] > 0$. Unless otherwise specified, expectation is taken over randomized treatment assignment.*

Put another way, Remark 1 states that, if the expected bias for feature $k$ is not much *smaller* than the expected bias for feature $h$, then the difference in the estimator for feature $k$ and the estimator for feature $h$ — the test statistic in Equation 6 — will be positive in expectation. Note that the bias for feature $k$ can be non-zero, so long as it is not sufficiently smaller than the bias for feature $h$.

Importantly, even when point estimates are biased, the test statistic in Equation 6 can still retain its useful properties: we expect a positive test statistic for the point estimate of the largest feature. Remark 1 also applies when the point estimate bias is small, or similar, across effects; in these settings, $E[B(h) - B(k)]$ will be close to zero. Further, Remark 1 takes advantage of the asymmetry of the most impactful feature. If the bias $B(k)$ tends to be large and positive, then it is less likely that $E[B(h) - B(k)]$ is sufficiently large enough to be larger than $\tau(k) - \tau(h)$. Similarly, if $B(h)$ is generally negative, or not as large as $B(k)$, then $E[B(h) - B(k)]$ may even be negative.

Ultimately, bias in the point estimates can actually *help* identify the top feature. Bias is only harmful when the bias of the estimator of the largest effect is smaller than the bias of the other features, and smaller by a wide margin. This is unlikely in our setting, as we will discuss in Section 6.2. Furthermore, since bias is often scale-sensitive, bias in the lower-impact features is less likely to overcome the $\tau(k) - \tau(h)$ gap (i.e., a 10% error on a low impact feature has a much smaller magnitude than a 10% error on a high impact feature). Asymmetry helps identify the top-ranked feature, which is often what practitioners care most about.

# 4 Robustness to bias

Randomized experiments are designed to extract causal estimates. Unfortunately, a number of forces can corrupt experiments and bias the traditional point estimates in factorial designs. Fortunately, it may be feasible to identify the most impactful feature even amidst this bias. Section 4.1 discusses estimation under measurement error, and Section 4.2 discusses estimation in experiments with conditional — and thus potentially unobserved — outcomes.

## 4.1 Measurement Error

Even in randomized experiments, measurement error can introduce bias (Gillen et al. 2019). Consider a general case where units exposed to feature $k$ independently experience some random measurement

error $e_k \sim [\mu_k, \sigma^2]$. When $\mu_k$ is not constant — perhaps some treatments are more elaborate, and more prone to measurement error — then the size of the error will be correlated to treatment. In our setting, we will see measurement error in the form of rounding error: true, latent intern performance, the quantity of interest, is rounded to the nearest integer to fit a rating system.

If the impact of measurement error is uncorrelated to treatment, then all of the estimated effects will be biased by the same amount in expectation. If the impact of measurement error is correlated to treatment, then certain estimates will be biased more than others. Not only will the point estimates be biased absolutely, but relative to each other. This has consequences for a conventional factorial design.

**Lemma 1** *Denote with $e_k$ the measurement error of units exposed to feature $k$. Let $e_k \sim [\mu_k, \sigma^2]$ independently, where $\mu_k$ is the average measurement error for units exposed to feature $k$ and $\sigma^2$ is the variance. Then, in a full factorial design, $\hat{\tau}(k)$ has bias $\mu_k$.*

All proofs are available in the Appendix.

When $\mu_k$ is large, point estimates will be severely biased. Fortunately, there may not be consequences for identifying the most impactful feature.

**Remark 2** *WLOG, let $\tau(k)$ be the largest and $\tau(h)$ be the second-largest treatment effects. Since $\hat{\tau}(k)$ is biased by $\mu_k$, we can write*

$$E[\hat{\tau}(k) - \hat{\tau}(h)] = \tau(k) - \tau(h) + \mu_k - \mu_h.$$

*Note that $\tau(k) - \tau(h) > 0$, and if $\tau(k) - \tau(h) > \mu_h - \mu_k$, then $E[\hat{\tau}(k) - \hat{\tau}(h)] > 0$.*

Just like Remark 1, measurement error can disrupt point estimates without impacting critical properties of the test statistic in Equation 6. This argument can apply in a few important settings. First, measurement error may not be large enough, or may not be different enough across treatments, to result in a difference $\mu_h - \mu_k$ that is larger than $\tau(k) - \tau(h)$. Second, when the measurement error $\mu_k$ is large and positive, it is less likely that $\mu_h - \mu_k$ is large and positive as well. Importantly, the test statistic in Equation 6 will be positive in expectation unless the difference in measurement error is larger than the difference in treatment effects. We will discuss why this scenario is unlikely in Section 6.2.

## 4.2 Conditional Outcomes

Often researchers wish to assess the impact of a treatment combination on an outcome that is observed conditionally on an intermediate outcome. Similarly, researchers might face non-compliance or attrition: units that do not take up treatment, or drop out before the experiment is over.

In the internship experiment we study there is a crucial conditional outcome: if an intern accepts an offer or not. By definition, this outcome is conditional on actually receiving an offer. If an intern does not receive an offer, then the conditional outcome is not well-defined.

One method to address conditional outcomes is by considering the "censoring due to death" problem discussed by Rubin (2006) using principal stratification (Frangakis and Rubin 2002). The principal stratification strategy was initially proposed for cases with binary treatment assignment and binary treatment receipt. It was later extended to include settings with continuous treatment receipt (Jin and Rubin 2008, Mattei and Mealli 2007, Schwartz et al. 2011), while the treatment assignment remained binary; recent work addresses settings with continuous treatments and continuous post-treatment variables (Antonelli et al. 2023).

In the following sections, we introduce principal stratification for factorial designs, which involves the opposite situation of non-binary assignment and binary treatment receipt. In this case, treatment assignment is determined by a combination of binary features that constitute a treatment allocation. However, the receipt of a specific treatment allocation — did unit $i$ receive treatment allocation $\boldsymbol{z}_j$, or not — *is* binary. We start by considering the special case with only one factor ($K = 1$), showing that it collapses to the usual principal stratification setting with binary treatment assignment and binary treatment receipt; then, we extend to the case with multiple factors ($K > 1$).

#### 4.2.1 Case $K = 1$

Let $W_i \in \{0, 1\}$ be the variable describing the assignment of unit $i$ to a binary treatment and denote with $w_i$ its realizations; denote with $Y_i$ the outcome (i.e., accepted an extended offer, or not), and indicate with $S_i(w_i) \in \{D, L\}$ whether the unit 'died' (did not receive an offer) or 'lived' (received an offer) after the experiment. Rubin (2006) defined four principal strata representing four different types of people: never survivors, $\{i : S_i(0) = D, S_i(1) = D\}$; always survivors, $\{i : S_i(0) = L, S_i(1) = L\}$; defiant survivors, $\{i : S_i(0) = L, S_i(1) = D\}$; compliant survivors, $\{i : S_i(0) = D, S_i(1) = L\}$.[2] A well-defined average causal effect of the active treatment exists only for the $LL$ group and can be denoted as the Survival Average Causal Effect (SACE), $E[Y_i(1) - Y_i(0)|S_i(0) = L, S_i(1) = L]$.

As in Rubin (2006), the internship experiment can be seen as a "censoring due to death" problem where 'death' is not receiving an offer. In our setting, SACE would be estimating the impact of treatment on acceptance rate among interns who would have always been extended an offer (no matter the treatment). A case where $K = 1$ (only one factor) implies the existence of $2^1 = 2$ treatment allocations, $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, when the factor is on or off. Therefore, this collapses to the standard binary assignment setting, as we can define $w_i = 1$ in place of $\boldsymbol{z}_1$ when the factor is on, and $w_i = 0$ in place of $\boldsymbol{z}_2$ when the factor is off. Further, we can write the compliance status of intern $i$ as $G_i = (S_i(0), S_i(1)) \in \{(D, D), (L, D), (D, L), (L, L)\}$ to denote never survivors, defiant survivors, compliant survivors and always survivors, respectively.

[2] In other words, never survivors and always survivors are those patients that, respectively, would not survive or always survive regardless of the assignment; defiant survivors would die under the active treatment and would survive under control; compliant survivors would survive under the active treatment and would die under control.

### 4.2.2 Case $K > 1$

In full factorial designs with two or more factors, we know from previous sections that there are $J = 2^K$ treatment allocations. For ease of notation, let $W_{i,j} = W_i(\boldsymbol{z}_j)$, with $W_{i,j} \in \{0, 1\}$, indicate if unit $i$ received treatment combination $\boldsymbol{z}_j$ and let $S_{i,j} = S_{i,j}(w_{i,j})$, with $S_{i,j} \in \{D, L\}$, be the status of unit $i$ after receiving treatment allocation $\boldsymbol{z}_j$.

To address full factorial designs with multiple treatments, we need to generalize our definition of compliant and defiant survivors.

**Definition 1** $G_{i,j} = (D, L)$*, or intern* $i$ *is a compliant survivor* with respect to treatment allocation $\boldsymbol{z}_j$*, if the intern would receive an offer under treatment allocation* $\boldsymbol{z}_j$ *and would* not *receive an offer under at least one treatment allocation in* $\boldsymbol{z}_{-j}$*.*

**Definition 2** $G_{i,j} = (L, D)$*, or intern* $i$ *is a defiant survivor* with respect to treatment allocation $\boldsymbol{z}_j$*, if the intern would* not *receive an offer under treatment allocation* $\boldsymbol{z}_j$ *and* would *receive an offer under at least one treatment allocation in* $\boldsymbol{z}_{-j}$*.*

In the notation of principal stratification, compliant survivors for the treatment allocation $\boldsymbol{z}_j$ are the set of interns $\{i : \exists S_{i,-j} = D, S_{i,j} = L\}$ and defiant survivors the set of interns $\{i : \exists S_{i,-j} = L, S_{i,j} = D\}$. The definitions for always survivors (never survivors) remain the same: an intern that would receive an offer (never receive an offer) under every treatment allocation. These can be written as $\{i : S_{i,j} = L \ \forall j\}$ and $\{i : S_{i,j} = D \ \forall j\}$.

It is important to point out that the compliance status is unobserved. For example, when we observe that an intern assigned to treatment $\boldsymbol{z}_j$ does not receive an offer, it could be because she is a defiant survivor with respect to treatment $\boldsymbol{z}_j$ or a never survivor and, without additional assumptions, it is not possible to identify this status from the data. Table 4 summarizes the situation. Table 5 illustrates a simple $2^2$ full factorial design with noncompliance.

Principal strata.

| $W_{i,j}$ | $S_{i,j}$ | $G_{i,j}$ |
|---|---|---|
| 1 | D | $(D, D) + (L, D)$ |
| 1 | L | $(L, L) + (D, L)$ |
| 0 | D | $(D, D) + (D, L)$ |
| 0 | L | $(L, L) + (L, D)$ |

Table 4: Each stratum is connected with the observed treatment assignment and treatment receipt.

### 4.2.3 Principal Stratification Estimation

In a setting with $K$ treatments, the survival average causal effect of treatment $k$, or $\text{SACE}_k$, can be written as

$$\text{SACE}_k = \frac{1}{2^{(K-1)}} \boldsymbol{g}_k' \left[ \bar{Y}(\boldsymbol{z}) | \boldsymbol{G}_i = (L, L) \right]. \tag{11}$$

$2^2$ full factorial design with noncompliance.

| $\boldsymbol{z}_j$ | Treatment 1 | Treatment 2 | $W_{i,j}$ | $S_{i,j}$ | Average potential outcomes | Compliance status |
|---|---|---|---|---|---|---|
| $\boldsymbol{z}_1$ | 1 | 1 | 1 | L | $\bar{Y}(\boldsymbol{z}_1)\|W_{i,1}=1, S_{i,1}=L$ | $(L,L)+(D,L)$ |
| $\boldsymbol{z}_1$ | 1 | 1 | 1 | D | $\bar{Y}(\boldsymbol{z}_1)\|W_{i,1}=1, S_{i,1}=D$ | $(D,D)+(L,D)$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\boldsymbol{z}_4$ | -1 | -1 | 1 | L | $\bar{Y}(\boldsymbol{z}_4)\|W_{i,4}=1, S_{i,4}=L$ | $(L,L)+(D,L)$ |
| $\boldsymbol{z}_4$ | -1 | -1 | 1 | D | $\bar{Y}(\boldsymbol{z}_4)\|W_{i,4}=1, S_{i,4}=D$ | $(D,D)+(L,D)$ |
| $\boldsymbol{z}$ | $\boldsymbol{g}_1$ | $\boldsymbol{g}_2$ | $\boldsymbol{w}_i$ | $\boldsymbol{S}_i$ | $\bar{Y}(\boldsymbol{z})\|W_i(\boldsymbol{z}), \boldsymbol{S}_i$ | $\boldsymbol{G}_i$ |

Table 5: In this table, $\bar{Y}(\boldsymbol{z}_1)|W_{i,1}=1, S_{i,1}=L$ denotes the average potential outcome across all units that are assigned to treatment $\boldsymbol{z}_1$ and receive an offer (survive).

As we can see from Table 5, all potential outcomes are mixtures. Under some additional assumptions, it can be possible to identify the compliance status and, in turn, estimate $\text{SACE}_k$ from the data. For instance, covariates can be used to impute the missing compliance status. In randomized experiments, if one can assume monotonicity (there are no defiant survivors) and that the proportion of observed groups is consistent across mixed groups, then the compliance status can similarly be imputed[3]. However, our setting lacks sufficient covariates, and we aim to develop a general method that does not require the assumptions of monotonicity and consistent proportions, especially because these assumptions can be more complicated in settings with multiple treatment allocations. Instead, we start by finding an estimator for the quantity defined in Equation 11. To do this, in principle we should consider the mixtures defined in the rows of Table 4 containing always survivors, i.e., second and fourth rows. However, notice that among the interns satisfying the conditions in the fourth row there are also defiant survivors: those that receive an offer if not assigned to $\boldsymbol{z}_j$ and do not receive an offer if assigned to $\boldsymbol{z}_j$, which means that we do not observe their outcome (accepting or declining an offer). This is analogous to the situation in Rubin (2006): one cannot observe the quality of life for people who died. Stated differently, even though defiant survivors and never survivors can *exist* in this experiment, we only *observe* outcomes for interns that survived (i.e., received an offer and made an acceptance decision). This implies that to define an estimator for $\text{SACE}_k$, we shall only consider the mixture composed by the always survivors and the interns who are compliant survivors with respect to the treatment they received. This leads to the following estimator for $\text{SACE}_k$:

$$\widehat{\text{SACE}}_k = \frac{1}{2^{(K-1)}} \boldsymbol{g}_k' \left[\bar{Y}^{obs}(\boldsymbol{z})|W_i(\boldsymbol{z})=1, \boldsymbol{S}_i=L\right] \tag{12}$$

$$= \frac{1}{2^{(K-1)}} \boldsymbol{g}_k' \left[(\boldsymbol{\pi}_{LL} \odot \bar{Y}^{obs}(\boldsymbol{z})|\boldsymbol{G}_i=(L,L)) + (\boldsymbol{\pi}_{DL} \odot \bar{Y}^{obs}(\boldsymbol{z})|\boldsymbol{G}_i=(D,L))\right], \tag{13}$$

where $\boldsymbol{\pi}_{LL}$ ($\boldsymbol{\pi}_{LD}$) is a $K \times 1$ column vector of probabilities, such that the $j^{th}$ entry indicates

[3]For example, under a monotonicity assumption, when $W_{i,j}=0$ and $S_{i,j}=L$ we can only observe always survivors and when $W_{i,j}=1$ and $S_{i,j}=D$ we can only observe never survivors, so if an equal number of never survivors and always survivors are observed, then the mixture group such that $W_{i,j}=1$ and $S_{i,j}=L$ is half compliers and half always survivors.

the probability that an intern who received treatment allocation $\boldsymbol{z}_j$ and received an offer is an always survivor (compliant survivor with respect to treatment allocation $\boldsymbol{z}_j$), and where $A \odot B$ denotes the Hadamard product of matrices $A$ and $B$. Here we condition on $W_i(\boldsymbol{z}) = 1, \boldsymbol{S}_i = L$, or intern $i$ getting an offer under the treatment allocation they received; this implies that the intern is an always survivor, or a compliant survivor relative to the treatment they received. Note that $\bar{Y}^{obs}(\boldsymbol{z})|W_i(\boldsymbol{z}) = 1, \boldsymbol{S}_i = L$ is the vector $(\bar{Y}^{obs}(\boldsymbol{z}_1)|W_{i,1} = 1, S_{i,1} = L, ..., \bar{Y}^{obs}(\boldsymbol{z}_K)|W_{i,K} = 1, S_{i,K} = L)$, i.e., the vector of average outcomes among interns who actually survive under the treatment allocations they receive.

Once again, the only well-defined effect, and the critical outcome that we wish to estimate, is $\frac{1}{2^{(K-1)}}\boldsymbol{g}_k'(\bar{Y}(\boldsymbol{z})|\boldsymbol{G}_i = (L, L))$, or the impact of an intervention in Week $k$ on acceptance rate among interns who would always be extended an offer regardless of which treatment allocation they received. Unfortunately, without additional assumptions or covariates to impute the compliance status, we cannot separate this effect from the acceptance rate among compliant survivors for each treatment allocation $\boldsymbol{z}_j$; that is, interns who would receive an offer under treatment allocation $\boldsymbol{z}_j$, but would not receive an offer under at least one other treatment allocation $\boldsymbol{z}_{-j}$. This threatens to bias our estimate.

**Lemma 2** *The estimator* $\widehat{SACE}_k$ *is biased by*

$$\frac{1}{2^{(K-1)}}\boldsymbol{g}_k'((\boldsymbol{\pi}_{LL} - 1) \odot \bar{Y}(\boldsymbol{z})|\boldsymbol{G}_i = (L, L)) + \frac{1}{2^{(K-1)}}\boldsymbol{g}_k'((1 - \boldsymbol{\pi}_{LL}) \odot \bar{Y}(\boldsymbol{z})|\boldsymbol{G}_i = (D, L)).$$

When all entries in $\boldsymbol{\pi}_{LL}$ are 0, the bias for treatment $k$ simply consists of the differences between complier outcomes and always survivor outcomes across treatment allocations. This is intuitive, since in this setting there are no always survivors, and we are only measuring outcomes for compliant survivors. In general, when entries of $\boldsymbol{\pi}_{LL}$ are between 0 and 1, the bias is a weighted difference of always survivor and compliant survivor outcomes.

There are a few settings where the bias in Lemma 2 will be zero. First, the estimator is unbiased if the average outcomes $\bar{Y}(\boldsymbol{z})$ are the same for always survivors and compliant survivors across all treatment allocations. Second, the estimator is unbiased if there are no compliant survivors in the experiment. Indeed, if all of the entries of $\boldsymbol{\pi}_{LL}$ are very close to 1, then the bias will be small.

However, it is likely that, in at least one treatment group, compliant survivors will both exist and have different average outcomes than always survivors. Perhaps always survivors are in general higher performing interns and have more outside options to choose from. Similarly, perhaps compliant survivors were on the margin of being extended an offer and feel fortunate enough to be nudged toward accepting. Any difference in average outcomes (in this experiment, the probability of accepting an offer) in any treatment allocation threatens bias in the point estimates. Fortunately, our approach may still maintain its important properties.

**Remark 3** *WLOG, let $\tau(k)$ be the largest and $\tau(h)$ be the second-largest treatment effects. Since $\hat{\tau}(k)$ is biased by $\delta_k = \frac{1}{2^{(K-1)}}\boldsymbol{g}_k'((\boldsymbol{\pi}_{LL} - 1) \odot \bar{Y}(\boldsymbol{z})|\boldsymbol{G}_i = (L, L)) + \frac{1}{2^{(K-1)}}\boldsymbol{g}_k'((1 - \boldsymbol{\pi}_{LL}) \odot \bar{Y}(\boldsymbol{z})|\boldsymbol{G}_i = (D, L))$, we can write*

$$E[\hat{\tau}(k) - \hat{\tau}(h)] = \tau(k) - \tau(h) + \delta_k - \delta_h.$$

*Note that $\tau(k) - \tau(h) > 0$, and if $\tau(k) - \tau(h) > \delta_h - \delta_k$, then $E[\hat{\tau}(k) - \hat{\tau}(h)] > 0$.*

Just like Remarks 1 and 2, bias can disrupt point estimates without affecting identification of the top feature. This result can apply in a few important settings. First, this result can hold when the difference in outcomes between the compliant survivor and always survivor groups is not large. Second, when the difference in outcomes between the compliant survivor and always survivor groups results in a large and positive bias for $\tau(k)$, which in turn makes $\delta_h - \delta_k$ less likely to be large and positive.

Importantly, the same approach may be applied to the problem of attrition in randomized experiments. Here, one would define 'surviving' as completing the study, and 'dying' as dropping out of the study. No interns dropped out of this experiment, but these results can be easily extended to situations where attrition did occur.

# 5 Simulation study

To test the merits of our approach, we conduct two simulation studies that closely mimic the experimental data from Bojinov et al. (2021). As hoped, the approach performs well — in the sense of identifying the highest impact feature — on simulated data. The tests maintain their accuracy under perturbations to important experimental variables (i.e., sample size). Our first simulation in Section 5.1 introduces measurement error, and our second simulation in Section 5.2 examines a setting with conditional outcomes. In both simulations, we are concerned with our ability to identify the most impactful feature.

## 5.1 Simulation: Intern Ratings

In this simulation study, we test if our method can reliably identify the week of intervention that has the largest impact on intern rating. The framework of our simulation is described in Section 5.1.1, and the results are given in Section 5.1.2.

### 5.1.1 Design

We generate data intended to closely imitate the experimental parameters described in Section 6.1: 737 'interns' exposed to $K = 4$ two-level features and $2^4 = 16$ feature combinations. Each feature $k$ denotes if the intervention — a water-cooler chat with senior management — was applied in Week $k$,

or not. The intervention can be applied in any combination (all, none, or some) across four weeks of the internship (Weeks 2-5), and each combination of weeks constitutes a single treatment allocation in a full factorial design.

For each of the $10,000$ iterations, we generate a $737 \times 16$ binary assignment matrix $W$. This determines which data is observed (1) and unobserved (0): if the value in row $i$, column $j$ is equal to 1, then intern $i$ received treatment $j$. Specifically, 737 entries in $W$ are randomly selected to contain 1, subject to the restriction that the rows of $W$ sum to 1 (each intern can be assigned to, at most, one treatment combination) and the columns of $W$ sum to the sample size in a given treatment group. To best represent the data in our example, we use the exact, unbalanced sample sizes in the experiment as the column sums of $W$. The allocation of these sample sizes to different treatment combinations is randomized at each iteration, except for the control group, which always has $n = 536$ interns to match the experiment.

For a single iteration, we generate the treatment effects $\boldsymbol{\tau}$ which, in our context, are defined as the average effect of intervening in a given week on an intern's rating. The effects in $\boldsymbol{\tau}$ can either be zero, small, moderate or large ($0, 0.1, 0.2, 0.3$, respectively). For instance, one simulation could consist of three treatments with small effects and one treatment with a large effect. We then generate the outcome for each intern, depending on which treatment combination they receive. For example, if the value at row $i$ and column $j$ of $W$ is 1, then we generate intern $i$'s outcome as 2 (an outcome in the middle of the rating scale) plus the first-order effects in $\boldsymbol{\tau}$ that are 'on' in treatment allocation $j$ (we assume higher-order effects are negligible). Finally, we add independent noise $N(0, \frac{1}{2})$ for each outcome to represent idiosyncratic randomness across interns.

The resulting data is continuous and represents the true, latent *performances* of each intern. We discretize the data by rounding the outcome to the nearest integer, ensuring a floor of 1 and a ceiling of 3. This provides simulated *ratings* — on a discrete scale from 1 to 3 — for the interns. Finally, to test our approach under different settings, we repeat the same simulation but vary important experimental parameters:

- Treatment effects: as described above, the treatment effects in $\boldsymbol{\tau}$ are assigned different sizes. We run a single simulation ($10,000$ iterations) for each of the following arrangements of $\boldsymbol{\tau}$: $(0, 0, 0, 0.1)$, $(0, 0, 0, 0.2)$, $(0, 0, 0, 0.3)$, $(0.1, 0.1, 0.1, 0.2)$, $(0.1, 0.1, 0.1, 0.3)$, $(0, 0, 0.1, 0.2)$, and $(0, 0, 0.1, 0.3)$.

- Number of Features: We increase the number of features $K$, doubling the sample size of 737 with each increment, since there are twice as many treatment combinations.

- Sample Size: We decrease the sample size across all treatment combinations by a fixed percentage $\gamma$.

- Noise: We increase the variance of the idiosyncratic noise used to generate each intern's rating.

- Continuous data: We do not round continuous intern performance to discrete intern rating in the final step.

#### 5.1.2 Results

To obtain insights from these simulations, the estimates defined in Equations 3 and 4 can be recovered using the simulated data. We then leverage the Fisher randomization test from Section 3.2 to identify the highest impact feature, and measure how often the test is correct. That is, if interventions in Week $k$ have the greatest effect on the outcome, we compute the rate of iterations that successfully estimate $\hat{R}_k = K$. We write this rate as $\theta^{(1)}$.

We use the following parameters as a baseline: three zero and one moderate treatment effect, $\sigma = 0.25$ noise of the outcome distribution, $K = 4$ features, $\gamma = 0$ decay rate. In this setting, $\theta^{(1)} = .93$. Therefore, we have high confidence in the ability of the test to identify the most impactful feature. The rate improves even further (99.6%) when using continuous data instead of rounding to discrete ratings.

Importantly, the point estimates in this simulation underestimate the true effect of the most impactful feature (0.20) by 0.09 on average. Clearly, the point estimates are severely biased — on average, the estimated impact of the top feature was just over half of its *true* impact — but our approach still identifies the top feature at a very high rate. This example demonstrates the merit of our approach: despite unreliable point estimates, the most impactful feature can be reliably extracted.

Table 6 depicts the critical measure — rate of identifying the top feature, or $\theta^{(1)}$ — when experimental parameters are varied away from the baseline parameters described above. Most of the arrangements of the treatment effects have high rates of detecting the single most impactful treatment. We find that $\theta^{(1)}$ falls when sample size is reduced or when the idiosyncratic noise is increased. However, $\theta^{(1)}$ increases with higher $K$, which gives us confidence in this method with more extensive experiments.

#### 5.1.3 Confidence Interval

It is challenging to find a theoretical confidence interval for this setting. Fortunately, we can use our simulation study — which closely mimics the data described in Section 6.1 — to find the empirical coverage of our test. The resulting interval tells us what effect sizes would be reliably detected.

We carry out the same default simulation from Section 5.1.1, but set the first three values in $\boldsymbol{\tau}$ to zero and vary the final (largest) treatment effect. Again, for each simulation ($10,000$ iterations), we calculate the rate of recovering the most impactful feature ($\hat{R}_k = K$ if interventions in Week $k$ have the greatest effect on the outcome). The complement of this rate gives us our p-value: the probability that the largest impact feature *is not* identified (i.e., is estimated as less than or equal to at least one other feature).

| Effect Sizes ($\boldsymbol{\tau}$) | Rate correct ($\theta^{(1)}$) |
|---|---|
| (0, 0, 0, 0.1) | 0.63 |
| *(0, 0, 0, 0.2)* | *0.93* |
| (0, 0, 0, 0.3) | 1.00 |
| (0.1, 0.1, 0.1, 0.2) | 0.74 |
| (0.1, 0.1, 0.1, 0.3) | 0.98 |
| (0, 0, 0.1, 0.2) | 0.83 |
| (0, 0, 0.1, 0.3) | 0.99 |
| Std. Dev ($\sigma$) | |
| 0.35 | 0.92 |
| 0.45 | 0.88 |
| 0.55 | 0.82 |
| 0.65 | 0.76 |
| 0.75 | 0.71 |
| 0.85 | 0.65 |
| 0.95 | 0.59 |

| No. of Features ($K$) | |
|---|---|
| 5 | 0.99 |
| 6 | 1.00 |
| 7 | 1.00 |
| 8 | 1.00 |
| Decay Rate ($\gamma$) | |
| 0.05 | 0.92 |
| 0.10 | 0.91 |
| 0.15 | 0.91 |
| 0.20 | 0.89 |
| 0.25 | 0.87 |
| 0.30 | 0.86 |
| 0.35 | 0.84 |
| 0.40 | 0.83 |

Table 6: $\theta^{(1)}$ is the rate of correctly identifying the most impactful feature. Each intern's outcome is simulated as 2 plus the sum of their treatment effects, plus zero-mean noise with variance $\sigma^2$. Each simulation has $K$ features and $\gamma$ sample size decay. The baseline simulation depicted in the second row is three zero and one moderate treatment effects, $\sigma = 0.25$, $K = 4$ and $\gamma = 0$; these parameters are held constant throughout, except for the adjusted parameter indicated in the first column.

The region of rejection, or the values of the largest treatment effect in $\boldsymbol{\tau}$ that can be identified at the 10% level of significance, constitute the empirical coverage of our proposed confidence interval. We find an interval of $[0.185, \infty)$. This concludes that our method would identify the most impactful feature at a 90% rate or better if that feature had an effect at least 0.185 larger than the rest of the features when $K = 4$. Note that this approach assumes that the features other than the most impactful feature all have an effect of zero. In reality, some features may have negative effects, which should make the identification of the largest feature even easier, and the lower bound of the coverage region even closer to zero.

## 5.2 Simulation: Intern Offers

In this simulation study, we test whether our method can reliably identify the week of intervention that has the largest impact on the probability that an intern accepts an extended offer. The framework of our simulation is described in Section 5.2.1, and the results are given in Section 5.2.2. Again, in practice, the other values of $\boldsymbol{\tau}$ may even be negative instead of zero. This would make the identification of the largest feature even easier.

### 5.2.1 Design

We generate data, the assignment matrix, the treatment effects $\boldsymbol{\tau}$ and the outcomes in the same fashion as Section 5.1.1. Since the outcome is a probability bounded between from 0 to 1, and not

a rating from 1 to 3, we adjust what we consider to be small, moderate, or large treatment effects (halved to $0.05, 0.10$ and $0.15$, respectively) and the standard deviation of the independent noise ($\frac{1}{8}$ instead of $\frac{1}{4}$). We generate the outcome in the same way as Section 5.1.1: summing the first-order effects in $\boldsymbol{\tau}$ that are 'on' in treatment allocation $j$, and adding the random noise. However, we then apply $\Phi$, the CDF of a $N(0,1)$ random variable, to ensure that the outcome is a probability. For example, if $\boldsymbol{\tau} = (0.2, -0.1, 0.3)$, then an intern in a treatment combination who received the first and second feature could generate outcomes with $\Phi(0.2 - 0.1 + \epsilon)$, where $\epsilon$ is the independent noise.

We designate 20% of the units in each treatment allocation to be compliers; again, we are interested in recovering the treatment effects for always survivors, and the existence of compliers complicates this estimation. The vast majority (91.5%) of the interns in the experiment received an offer, which means that the vast majority of the interns were always survivors or compliers. Since it was rare for interns to not receive an offer, we believe that compliers — interns who did receive an offer, but would not have received an offer under at least one other treatment allocation — are not common, and that 20% is a reasonable amount of compliers. To allow the treatment effects to differ between compliers and always survivors, we draw the complier values of $\boldsymbol{\tau}$ independently from a Unif$(-.15, .15)$ distribution. This represents a situation where the effects for always survivors are completely unrelated to the effects for compliers, which makes estimation challenging.

To test our approach under different settings, we repeat the same default simulation but vary the important experimental parameters from Section 5.1.1: values of $\boldsymbol{\tau}$, number of features, sample size, and noise. We add one further variation:

- Complier rate: we increase the number of compliers in each treatment allocation.

### 5.2.2 Results

Once again, the estimates defined in Equations 3 and 4 can be recovered using the simulated data, and we evaluate our approach in terms of its ability to identify the highest impact feature. Once again, we write the rate of identifying the highest impact feature as $\theta^{(1)}$ ($\hat{R}_k = K$). In the default setting of $\sigma = 0.125, K = 4, \gamma = 0$, three zero effects and one moderate effect, and 20% of each group across treatment allocations being compliers, the rate of detection of the most impactful feature is very high: $\theta^{(1)} = 0.96$.

Table 7 depicts the critical measure, the rate of identification of the top characteristic, or $\theta^{(1)}$, when the experimental parameters are varied away from the baseline parameters described above. We see similar results as Section 5.1.2: lower $\theta^{(1)}$ for higher noise, lower sample size and a larger complier rate. Once again, $\theta^{(1)}$ climbs with $K$.

### 5.2.3 Confidence Interval

We use the default simulation study from Section 5.2.1 to estimate the empirical coverage of our proposed confidence interval. Again, we set the first three values in $\boldsymbol{\tau}$ to zero and adjust the final

| Effect Sizes ($\boldsymbol{\tau}$) | Rate correct ($\theta^{(1)}$) |
|---|---|
| (0, 0, 0, 0.05) | 0.71 |
| *(0, 0, 0, 0.2)* | *0.96* |
| (0, 0, 0, 0.15) | 1.00 |
| (0.05, 0.05, 0.05, 0.1) | 0.70 |
| (0.05, 0.05, 0.05, 0.15) | 0.96 |
| (0, 0, 0.05, 0.1) | 0.84 |
| (0, 0, 0.05, 0.15) | 0.98 |
| Std. Dev ($\sigma$) | |
| 0.100 | 0.98 |
| 0.150 | 0.94 |
| 0.175 | 0.90 |
| 0.200 | 0.87 |
| 0.225 | 0.83 |
| 0.250 | 0.78 |
| 0.275 | 0.75 |
| 0.300 | 0.72 |
| No. of Features ($K$) | |
| 5 | 0.99 |
| 6 | 1.00 |
| 7 | 1.00 |
| 8 | 1.00 |

| Decay Rate ($\gamma$) | |
|---|---|
| 0.05 | 0.96 |
| 0.10 | 0.95 |
| 0.15 | 0.94 |
| 0.20 | 0.94 |
| 0.25 | 0.93 |
| 0.30 | 0.93 |
| 0.35 | 0.92 |
| 0.40 | 0.92 |
| Complier Rate ($\nu$) | |
| 0.10 | 1.00 |
| 0.15 | 0.98 |
| 0.25 | 0.91 |
| 0.30 | 0.84 |
| 0.35 | 0.77 |
| 0.40 | 0.68 |
| 0.45 | 0.61 |
| 0.50 | 0.56 |

Table 7: $\theta^{(1)}$ is the rate of correctly identifying the most impactful feature. Each intern's outcome is simulated as 2 plus the sum of their treatment effects, plus zero-mean noise with variance $\sigma^2$. Each simulation has $K$ features and $\gamma$ sample size decay. The proportion of compliers in each treatment allocation is given by $\nu$. The baseline simulation depicted in the second row is three zero and one moderate treatment effects, $\sigma = 0.125$, $K = 4, \gamma = 0$ and $\nu = 0.20$; these parameters are held constant throughout, except for the adjusted parameter indicated in the first column.

(largest) value, and for each simulation (10,000 iterations), we calculate the rate of recovery of the most impactful feature. As before, the complement of this rate gives us our p-value.

We find an interval of $[0.082, \infty)$. This concludes that our method would identify the most impactful feature at a 90% rate or better if the largest value of $\boldsymbol{\tau}$ was at least 0.082 greater than the next largest value of $\boldsymbol{\tau}$ when $K = 4$. Again, this approach assumes that the features other than the most impactful feature have an impact of zero; in practice, secondary features with negative effects can make identifying the most impactful feature easier.

# 6 Empirical Applications

To illustrate the benefit of our approach, we re-analyze part of an experiment run at a large financial institution (Bojinov et al. 2021). In particular, we focus on this experiment because the outcome is subject to measurement (rounding) error and there is a secondary conditional outcome.

## 6.1 Data

The data consists of $n = 1,370$ interns completing a summer internship at a leading financial institution (Bojinov et al. 2021). The interns worked remotely across 16 cities. The gender split is $45\% - 55\%$, although we do not know in which direction (the gender values are anonymized).

The interns were randomly and independently assigned to one of four active treatment groups or a control, following a standard panel experiment structure (Bojinov and Shephard 2019, Bojinov et al. 2019, Han et al. 2024). The active treatments were an asynchronous Q/A, an intern group project, an intern-only virtual water cooler chat or a virtual water cooler chat with a senior manager. We will focus our analysis on the senior manager chat group, which we simply call the *treatment group*; this was the only active treatment in Bojinov et al. (2021) to represent all $2^4$ treatment combinations across weeks (although it is unbalanced) *and* suggest an actual improvement in the results. For interns assigned to the treatment group, the chat could take place in Weeks 2, 3, 4, and/or 5 of the internship. We consider different intervention timings as different treatments, and thus have $2^4$ possible treatment combinations (e.g., an intern can have a chat in Weeks 2 and 3, in Week 4 only, in all weeks, in none of the weeks, etc.). In this paradigm, the experiment can be modeled as a full factorial design. We aim to understand if the timing of treatment moderated its positive impact.

The primary outcome that we study is the performance rating at the intern level.[4] Each week, the intern's manager enters a rating that reflects the intern's performance; in general, performance ratings in this style are critical processes in business operations (Cappelli and Conyon 2018). In this experiment, each rating is given on an integral scale of 1 to 3, with 3 being the best score. Managers are not allowed to enter any rating other than $1, 2$ or 3: the resulting imprecision makes it challenging

[4]There are a number of other practically relevant outcomes in this data, including survey responses from the interns discussing important facets of the internship. These are interesting areas of future study.

to understand the impacts of interventions on true intern performance at a more granular level. In other words, we can use these data to extract an unbiased estimate of the intern *rating*, but we care more about true, latent intern *performance*. The integral scale that managers use is coarse: there are only two gradations, so it is difficult to fully distinguish between interns. It is more useful to model the intern's true, precise performance on a continuous scale from 1 to 3, with the manager rounding this performance to the nearest integer.

Ultimately, we expect nearly every rating to be separated by rounding error from an intern's true, latent rating. For example, an intern with a 'true' rating of 2.341 might be rounded to 2, a much less descriptive score. Indeed, ratings of 3 have almost always been rounded up, and ratings of 1 have almost always been rounded down. This has consequences for estimating the impact of timing on true, latent intern performance.

After the internship is completed, the interns can receive an offer or not; if they receive an offer, they can accept the offer or not. Therefore, our secondary outcome is the acceptance rate of the offer, as retention is a critical component of corporate strategy (Cloutier et al. 2015). We only observe the acceptance decisions of interns who receive an offer, and thus we cannot know whether the interns who did receive an offer would have accepted a potential offer. This makes standard testing challenging (Frangakis and Rubin 2002). The causal effect of treatment is only well-defined for 'always survivors', or interns that would have received an offer regardless of their treatment status. Unfortunately, for a treated intern that has received an offer, we cannot determine whether they are an always survivor or a complier, as we do not observe if they would have received an offer if they were not treated.

Given the challenges outlined above, instead of focusing on point estimates of the causal effects of treatments in different weeks, we collect rank estimates of these effects and identify the most impactful feature. This approach is more robust to the sources of bias inherent in these data. Ultimately, the objective of our empirical analysis is to identify at what time interns should receive treatment to maximize their performance and the probability of accepting an offer that is extended to them.

## 6.2 Empirical Application: Intern Ratings

Almost every intern rating is separated from the true underlying intern performance due to rounding error. This is a special case of measurement error, which is conducive to the arguments in Remark 2. Specifically, while the point estimates are likely to be biased, identification of the most impactful feature may not be disrupted. There are $n = 737$ interns who were either assigned to the senior water cooler session or were assigned to the control group; we focus our analysis on these interns. [5]

[5] We exclude 23 interns who did not receive a final rating from the analysis and assume that these ratings are missing at random (Marini et al. 1980). All interns completed the internship, but in these 23 instances managers did not enter a final rating, because they forgot or missed the deadline. 'Missing' status is not strongly associated with nearly all of our covariates, and even under some worst-case assumptions our results are still significant; we discuss this analysis in the Appendix. Further, Bojinov et al. (2021) make a strong assumption that the data are missing completely at random.

| Outcome: Week 5 rating | | |
|---|---|---|
| Treatment Week | Effect on Rating | Rank |
| 5 | 0.199 | 4 |
| 2 | 0.179 | 3 |
| 3 | -0.126 | 2 |
| 4 | -0.134 | 1 |

Table 8: Estimated average effect of treatment during a specific week on the final intern rating.

We construct an unbalanced, full $2^4$ factorial design in which the treatment arms represent the combination of weeks that an intern received treatment. The estimated treatment effects on the intern's final (Week 5) performance rating are provided in Table 8. These effects are quite large: for example, an effect of 0.20 is 10% of the range from the minimum score (1) to the maximum score (3). However, because of the rounding error impacting each intern rating, we cannot have much confidence in the accuracy of the point estimates.

Fortunately, in settings like these, firms may not be interested in recovering exact point estimates. Instead, they may prefer to identify the week with the *largest* impact. After all, these interventions are costly because they require a time investment from senior management. Firms may want to conduct the treatment in just a single week of the internship, and thus their primary goal may be to identify the most impactful time to intervene (if there is one).

The arguments in Remark 2 will apply if the average rounding errors are not large, or different, enough to overcome the difference in the true treatment effects. We believe that at least one of these conditions will be satisfied. It is not likely that the rounding error across treatments would be both large and significantly different. For such a disruption to occur, we would either need ratings for interns in a lower-impact treatment to be rounded up while ratings for interns in a higher-impact treatment are rounded down, or not sufficiently rounded up; both are unlikely. First, rounding stops at the nearest integer, and thus has a ceiling of 0.5; we do not expect large differences in rounding error. Second, ratings in higher-impact treatments are more likely to be rounded up, since these ratings are a priori more likely to land in the 'round up' regions ($1.5-2$, $2.5-3$) than in the 'round down regions' ($1-1.5$, $2-2.5$), and vice versa for lower-impact treatments. This leverages the asymmetry central to our method: positive bias in treatments with the highest-impact feature actually *helps* in correctly identifying that feature.

In summary, our method is well-suited for this setting. The point estimates are expected to be biased, but identification of the most impactful feature is relatively robust to this bias — and may even be aided by bias. Further, the main concern of practitioners in this scenario is usually identifying the top feature — that is, the timing of the intervention that has the greatest impact on intern performance. The precise point estimates of the features, as well as the second-, third- and fourth-largest features, are of less practical importance.

Table 8 shows that $\hat{R}_5 = 4$ and $\hat{R}_2 = 3$. It is interesting that treatment at the start and end

| Outcome: Week 5 rating | | | | |
|---|---|---|---|---|
| Treatment Week | Effect | Rank | $H_0$ | p-value |
| 5 | 0.199 | 4 | $\tau_5 - \tau_2 = 0$ | 0.789 |
| 2 | 0.179 | 3 | $\tau_2 - \tau_3 = 0$ | 0.000*** |
| 3 | -0.126 | 2 | $\tau_3 - \tau_4 = 0$ | 0.912 |
| 4 | -0.134 | 1 | | |
| Outcome: Week 4 rating | | | | |
| Treatment Week | Effect | Rank | $H_0$ | p-value |
| 2 | 0.146 | 3 | $\tau_2 - \tau_3 = 0$ | 0.053* |
| 3 | -0.050 | 2 | $\tau_3 - \tau_4 = 0$ | 0.331 |
| 4 | -0.138 | 1 | | |
| Outcome: Week 3 rating | | | | |
| Treatment Week | Effect | Rank | $H_0$ | p-value |
| 2 | 0.111 | 2 | $\tau_2 - \tau_3 = 0$ | 0.613 |
| 3 | 0.050 | 1 | | |
| Outcome: Week 2 rating | | | | |
| Treatment Week | Effect | Rank | $H_0$ | p-value |
| 2 | 0.094 | 1 | $\tau_2 = 0$ | 0.081* |

Table 9: Each p-value indicates the significance of testing if the effect is greater than the next largest estimated effect: $^{*}\, p < 0.1$; $^{**}\, p < 0.05$; $^{***}\, p < 0.01$.

of the internship had the largest impact on final intern rating. One potential explanation is that treatment at the start of the internship has an enduring impact throughout the remainder of the internship, while treatment in the final week provides a boost in the week it was applied. To assess this, we test if the observed differences in point estimates are statistically significant for ratings in Weeks 2-5 using $10,000$ treatment re-assignments under Hypothesis 1 (the effects of a single week can also be tested against zero). This informs us if treatment impacts are immediate (i.e., treatment in Week 3 improving Week 3 ratings) and/or if early treatments have effects that endure over time (i.e., treatment in Week 2 improves Week 4 ratings). Note that some treatment allocations have sample sizes as small as $n = 6$, so we keep the number of interns the same within each treatment during the Fisher testing to reflect this imbalance.

Results are reported in Table 9. We find that intervening in Week 2 has a positive impact — both immediate and enduring — on intern performance. For example, Table 9 shows that a Week 2 intervention has a significantly larger effect than the next largest intervention on Week 4 ratings and final, Week 5 ratings. Following the confidence interval paradigm from Section 3.2, we find $c = 0.032$ when considering the effect of interventions in Weeks 2, 3 and 4 on the rating in Week 4. That is, at the 10% level of significance, the results indicate that an intervention in Week 2 has an effect at least 0.032 larger than an intervention in Week 3 on the Week 4 rating. When considering the impact on Week 5 ratings, we find $c = 0.182$, indicating that at the 10% level of significance, we conclude that an intervention in Week 2 has an effect at least 0.182 larger than an intervention in Week 3 on the final Week 5 rating.

| Outcome: Week 2 rating | | | | |
|---|---|---|---|---|
| Treatment Week | Effect | Rank | $H_0$ | p-value |
| 3 | 0.046 | 3 | $\tau_3 - \tau_5 = 0$ | 0.473 |
| 5 | -0.013 | 2 | $\tau_5 - \tau_4 = 0$ | 0.962 |
| 4 | -0.017 | 1 | | |
| Outcome: Week 3 rating | | | | |
| Treatment Week | Effect | Rank | $H_0$ | p-value |
| 5 | 0.073 | 2 | $\tau_5 - \tau_4 = 0$ | 0.343 |
| 4 | -0.044 | 1 | | |
| Outcome: Week 4 rating | | | | |
| Treatment Week | Effect | Rank | $H_0$ | p-value |
| 5 | 0.066 | 1 | $\tau_5 = 0$ | 0.211 |

Table 10: $^{*}\, p < 0.1$; $^{**}\, p < 0.05$; $^{***}\, p < 0.01$.

Further, Table 9 indicates that Week 5 interventions also have a large effect on the final (Week 5) rating, but not significantly larger than Week 2 interventions. This suggests a short-term increase in ratings from an intervention in the week it was administered. Indeed, Week 2 interventions have an effect significantly above zero on Week 2 ratings. Following the confidence interval paradigm, we find $c = 0.014$, so at the 10% level of significance, we conclude that an intervention in Week 2 has an effect size of at least 0.014 on the Week 2 rating.

Ultimately, there is evidence that the earliest intervention possible — Week 2 — has an instant impact on ratings in Week 2 and a stronger impact on future ratings (Weeks 4 and 5) than even interventions in later weeks (Weeks 3 and 4). Furthermore, there is suggestive, non-significant evidence that an intervention in Week 2 has a larger impact on Week 3 ratings than an intervention in Week 3 ($p = 0.613$). Overall, our results imply that interventions at the beginning of the internship have the greatest aggregate impact: a significant jump when they are deployed that then endures throughout the period.

Finally, we conduct simple placebo checks: do future interventions have a significant impact on previous week ratings? The answer, of course, should be no. Fortunately, the data shows no signs of spurious relationships. The results in Table 10 indicate that no intervention has a significant impact on a past rating.

## 6.3 Robustness: Intern Offers

Next, we estimate how the *timing* of an intervention affects the likelihood that an intern accepts an offer. This data is naturally censored: we only observe the outcome — decision to accept an extended offer, or not — for interns who actually receive an offer. As discussed in Section 4.2, these 'surviving' interns can be split into always survivors, who receive an offer regardless of treatment assignment, and compliant survivors, who receive an offer only if they receive the treatment at least once.

| Outcome: Offer Acceptance | | |
|---|---|---|
| Treatment Week | Effect on Acceptance | Rank |
| 2 | 6.0% | 4 |
| 3 | −1.5% | 3 |
| 5 | −4.9% | 2 |
| 4 | −11.8% | 1 |

Table 11: Estimated average effect of treatment timing on the probability of accepting an extended offer.

Our aim is to recover the treatment effect among always survivors. Practically, these might be the highest performing interns, since they did not need an intervention to receive an offer. More importantly, though, treatment effects are only well-defined for always survivors because their outcomes are defined under both treatment and control. In contrast, our causal effect of interest is not well-defined for compliant survivors, since there is no control counterfactual: there would be no offer to accept in the control case. We construct a full $2^4$ factorial design to estimate the impact of treatment in different weeks on the probability of accepting an extended offer among always survivors.

The results in Table 11 provide the estimates. For instance, we estimated that an intervention in Week 3 has an average effect of $-1.5\%$ on the probability of accepting an extended offer among always survivors. Unfortunately, in the absence of covariates that could impute intern compliance status or additional assumptions discussed in Section 4.2, it is nearly impossible to differentiate the treatment effect on always survivors from the treatment effect on compliant survivors. The bias cannot be considered negligible: while we can extract a bound using Lemma 2, the bound is not useful in a setting with a binary outcome. Consider a simple, extreme example. If all interns are compliant survivors, and the potential outcomes are 1 under treatment allocation $z_j$ (0 under all other treatment allocations $z_{-j}$), then estimates will be biased by $\frac{1}{32}(32) - 0 = 1$, since $K = 4$. Put more simply, we could be estimating the effect on compliant survivors instead of always survivors, and the difference between compliant and always survivors could be maximally 1 (always accept vs. never accept).

Fortunately, per Remark 3, the test statistic in Equation 6 can still retain its important properties. Although we do not expect the probability of acceptance to be equal across always survivors and compliant survivors, it is unlikely that the difference is large: among interns who were either assigned to the senior water cooler session or were assigned to the control group, the acceptance rate is quite high (84.8%). Further, since the vast majority of interns receive an offer (93.1%) it is unlikely that there are a large number of compliant survivors, or interns who would not have received an offer in at least one other treatment allocation than the allocation they received.

This setting represents another promising application of our method. We expect bias in the point estimates, but perhaps not a disruption in identifying the most impactful feature. Ultimately, though, the results of the Fisher randomization test using the paradigm of Hypothesis 1 — with

| Outcome: Offer Acceptance | | | | |
|---|---|---|---|---|
| Treatment Week | Effect | Rank | $H_0$ | p-value |
| 2 | 6.0% | 4 | $\tau_2 - \tau_3 = 0$ | 0.190 |
| 3 | −1.5% | 3 | $\tau_3 - \tau_5 = 0$ | 0.430 |
| 5 | −4.9% | 2 | $\tau_5 - \tau_4 = 0$ | 0.202 |
| 4 | −11.8% | 1 | | |

Table 12: Each p-value indicates the significance of testing if the effect is greater than the next largest estimated effect. $^{*}\, p < 0.1$; $^{**}\, p < 0.05$; $^{***}\, p < 0.01$.

10,000 treatment re-assignments — does not indicate that the results are significant. The results are portrayed in Table 12, and are suggestive that the earliest possible intervention — Week 2 — has the largest effect on accepting an extended offer. Indeed, this week has the only positive estimated effect. However, this effect is not significantly larger than the second-most impactful intervention in Week 3 ($p = .190$).

In this setting, where there is low variability in the outcome — the vast majority of interns who received an offer accepted that offer — it is challenging to establish statistical significance. In general, we hope that this example demonstrates an approach that can address common problems in the literature. When data are censored by conditional outcomes — and lack sufficient covariates to impute compliance status — a Fisher randomization test can extract important insights. The same approach can apply to classical attrition, where units may drop out of an experiment.

# 7 Conclusion

Factorial experiments, which allow researchers to test multiple treatments, each with multiple levels, have grown in popularity. Researchers and practitioners can leverage factorial designs in complex settings to decide between many possible interventions, each of which is costly to deploy. In these experiments, identifying the treatment that has the largest impact is more important than extracting the exact point estimate of every treatment. Fortunately, finding the most impactful treatment can still be possible under common sources of experimental bias. We defined the non-negligible bias in factorial design estimates under measurement error and data censoring and showed how the asymmetry of the largest treatment effect — positive bias *helps* identify this treatment — allows researchers to find the most impactful treatment even when point estimates are not trustworthy. We introduced a Fisher randomization test to perform non-parametric inference that identifies the largest treatment effect.

We demonstrated our approach using an existing field experiment at a large financial firm, where a costly intervention — water cooler chats with senior management — was deployed at different points during an internship. Important measures were obscured by measurement error and data censoring: intern performance ratings were coarsely assigned on a discrete scale from 1 to 3, and

offer acceptance decisions from interns who did not receive an offer could not be observed. We have little confidence in the point estimates, which are almost certainly biased. Fortunately, our inferential approach suggests that the earliest treatments were the most impactful. Water cooler chats in Week 2 — the earliest week that the chats were held — had large immediate and lasting effects on intern outcomes, including larger effects on ratings in future weeks than interventions in those future weeks. This result provides an important lesson across the diverse landscape of workplace programs: intervening early may have the largest impact.

Our study is not without limitations. This experiment focuses only on interns, and the outcomes measured are performance and retention. While our results are generalizable, it is possible that these insights do not apply in all contexts and may not hold for full-time employees or other outcomes that were not measured in the study. We found suggestive, but not significant, results for intern offer acceptance; since most interns accepted extended offers, the observed outcomes have little variation.

# A Proofs of Lemmas and Theorems

**Proof 1** *Proof of Lemma 1*

*From (Pashley and Bind 2023), we have the unbiased estimator for the average potential outcome for treatment $z_j$ in the case without measurement error:*

$$\bar{Y}^{obs}(\boldsymbol{z}_j) = \frac{1}{n_j}\sum_{i=1}^{N} W_i(z_j)Y_i(\boldsymbol{z}_j) = \frac{1}{n_j}\sum_{i:W_i(z_j)=1} Y_i^{obs}(z_j).$$

*Now consider the case of measurement error in feature $k$, which is introduced when feature $k$ is 'on'; that is, feature $k$ is included in treatment allocation $z_j$ (if feature $k$ is 'off', or not included in the treatment allocation, there is no impact of measurement error from feature $k$). Consider some other feature $h \neq k$. If the measurement error from feature $h$ is independent of other features included in a treatment allocation, then this measurement error will not impact the estimate of feature $k$. This is because feature $h$ is 'on' in exactly half, and 'off' in exactly half, of the treatment allocations where feature $k$ is 'on'; feature $h$ is also 'on' in exactly half, and 'off' in exactly half, of the treatment allocations where feature $k$ is 'off'. Ultimately, the expected impact of feature $h$ measurement error cancels out.*

*The estimator of the average potential outcome for some treatment $z_j$ where feature $k$ is on is given by*

$$\bar{Y}_e^{obs}(\boldsymbol{z}_j) = \frac{1}{n_j}\sum_{i:W_i(z_j)=1} (Y_i^{obs}(z_j) + e_{i,k}) \tag{14}$$

*where $e_{i,k} \sim [\mu_k, \sigma^2]$ is the measurement error for unit $i$ from feature $k$ exposure and $\bar{Y}^{obs}(z_j)$ is the estimator for the average potential outcome for treatment $z_j$ without measurement error. Taking expectations of both sides yields*

$$\begin{aligned} E[\bar{Y}_e^{obs}(\boldsymbol{z}_j)] &= E\left[\frac{1}{n_j}\sum_{i:W_i(z_j)=1} Y_i^{obs}(z_j)\right] + \frac{1}{n_j}\sum_{i:W_i(z_j)=1} E[e_{i,k}] \\ &= E[\bar{Y}^{obs}(z_j)] + \mu_k \\ &= \bar{Y}(z_j) + \mu_k. \end{aligned}$$

*The last equality comes from the unbiasedness of $\bar{Y}^{obs}(\boldsymbol{z}_j)$ in the absence of measurement error. Put together, the above derivation shows that, in the case of measurement error for feature $k$, the estimator $\bar{Y}_e^{obs}(\boldsymbol{z}_j)$ has bias $\mu_k$ for the average potential outcome under treatment $\boldsymbol{z}_j$.*

*Now let $\bar{\boldsymbol{Y}}_e^{obs}(\boldsymbol{z}) = (\bar{Y}_e^{obs}(\boldsymbol{z}_1), \ldots, \bar{Y}_e^{obs}(\boldsymbol{z}_J))$ indicate the vector of observed mean potential outcomes in the case of measurement error. The estimator for $\tau(k)$ in the presence of measurement error becomes*

$$\hat{\tau}(k) = \frac{1}{2^{K-1}}\boldsymbol{g}_k'\bar{\boldsymbol{Y}}_e^{obs}(\boldsymbol{z}).$$

*We have shown that* $E\left[\bar{Y}_e^{obs}(\boldsymbol{z}_j)\right] = \bar{Y}(\boldsymbol{z}_j) + \mu_k$ *for treatments where* $k$ *is on, and the same principle extends to* $\bar{\boldsymbol{Y}}_{\boldsymbol{e}}^{obs}(\boldsymbol{z})$*. Taking expectations of both sides:*

$$\begin{aligned} E\left[\hat{\tau}(k)\right] &= E\left[\frac{1}{2^{K-1}}\boldsymbol{g}_k'\bar{\boldsymbol{Y}}_e^{obs}(\boldsymbol{z})\right] \\ &= \frac{1}{2^{K-1}}\boldsymbol{g}_k' E\left[\bar{\boldsymbol{Y}}_e^{obs}(\boldsymbol{z})\right] \\ &= \underbrace{\frac{1}{2^{K-1}}\boldsymbol{g}_k'\bar{\boldsymbol{Y}}(\boldsymbol{z})}_{\tau(k)} + \frac{2^{K-1}\mu_k}{2^{K-1}} \end{aligned}$$

*as there are* $2^{K-1}$ *treatments for which feature* $k$ *is 'on' and measurement error is introduced, and these treatments coincide with the* 1 *entries of* $\boldsymbol{g}_k'$*. Thus, we are left with* $\hat{\tau}(k) = \tau(k) + \mu_k$*.* ■

**Proof 2** *Proof of Lemma 2 The bias is calculated as*

$$E[\widehat{SACE}_k - SACE_k] = E[\widehat{SACE}_k] - \frac{1}{2^{(K-1)}}\boldsymbol{g}_k'(\bar{Y}(\boldsymbol{z})|\boldsymbol{G}_i = (L,L)).$$

*Since* $\bar{Y}^{obs}(\boldsymbol{z})|\boldsymbol{G}_i = (L,L)$ *is unbiased for* $\bar{Y}(\boldsymbol{z})|\boldsymbol{G}_i = (L,L)$*, and* $\bar{Y}^{obs}(\boldsymbol{z})|\boldsymbol{G}_i = (D,L)$ *is unbiased for* $\bar{Y}(\boldsymbol{z})|\boldsymbol{G}_i = (D,L)$*, we can re-write* $E[\widehat{SACE}_k]$ *in the above expression:*

$$\frac{1}{2^{(K-1)}}\boldsymbol{g}_k'\left[(\boldsymbol{\pi}_{LL} \odot \bar{Y}(\boldsymbol{z})|\boldsymbol{G}_i = (L,L)) + (\boldsymbol{\pi}_{DL} \odot \bar{Y}(\boldsymbol{z})|\boldsymbol{G}_i = (D,L))\right] - \frac{1}{2^{(K-1)}}\boldsymbol{g}_k'(\bar{Y}(\boldsymbol{z})|\boldsymbol{G}_i = (L,L))$$

$$= \frac{1}{2^{(K-1)}}\boldsymbol{g}_k'((\boldsymbol{\pi}_{LL} - 1) \odot \bar{Y}(\boldsymbol{z})|\boldsymbol{G}_i = (L,L)) + \frac{1}{2^{(K-1)}}\boldsymbol{g}_k'((1 - \boldsymbol{\pi}_{LL}) \odot \bar{Y}(\boldsymbol{z})|\boldsymbol{G}_i = (D,L))$$

*since* $\boldsymbol{\pi}_{DL} = 1 - \boldsymbol{\pi}_{LL}$*.* ■

# B Missing at Random Assumption: Experiment Data

We discuss in Section 6.2 how 23 of our sub-sample of 737 interns did not receive a final rating and thus were dropped from the analysis. We justified this removal by assuming these 23 interns were missing at random; this is reflected by the 'missing completely at random' assumption in Bojinov et al. (2021). The following regression tables address that claim. There is no significant difference (at the 5% level) in gender, ethnicity, or Division between interns who received a final rating and interns who did not.

One significance to note is in receiving an offer. Interns who did not have a final rating did appear to be less likely to receive an offer; further, these interns had slightly lower Week 2 ratings

and slightly higher Week 4 ratings. It is our belief that this difference is due to a combination of small sample size, randomness, and rare events (not receiving an offer), but this is important to note.

Further, the only treatments that are significantly positively associated with not receiving a final rating are Treatment Arms 12 and 15. Both of these arms apply treatment in Week 5, and Treatment Arm 12 applies treatment in Week 2. We found Weeks 5 and 2 to have the most positive impact on final performance. Therefore, there is the potential that these estimated effects were inflated by poor-performing interns in these treatments dropping out of the internship.

Fortunately, even the worst-case scenario does not disrupt our results. If the three interns in Treatments 12 and 15 received a rating of 1 — in and of itself a rare event, as only 4% of interns received this rating — then results of the Fisher randomization would still be significant. The point estimates would drop down to 0.16 and 0.12 for Week 5 and Week 2, respectively.

Table 13

| | *Dependent variable:* |
|---|---|
| | dropped |
| Gender_Flag | −0.011 |
| | (0.013) |
| Ethnicity: Brown | −0.002 |
| | (0.133) |
| Ethnicity: Green | 0.041 |
| | (0.050) |
| Ethnicity: Indigo | −0.007 |
| | (0.181) |
| Ethnicity: Orange | 0.044 |
| | (0.052) |
| Ethnicity: Pink | −0.002 |
| | (0.062) |
| Ethnicity: Red | −0.002 |
| | (0.133) |
| Ethnicity: Violet | 0.025 |
| | (0.050) |
| Ethnicity: Yellow | 0.014 |
| | (0.051) |
| Constant | 0.007 |
| | (0.049) |
| Observations | 737 |
| $R^2$ | 0.007 |
| Adjusted $R^2$ | −0.006 |
| Residual Std. Error | 0.174 (df = 727) |
| F Statistic | 0.549 (df = 9; 727) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 14

| | *Dependent variable:* |
|---|---|
| | dropped |
| Offer_Made_Flag | −0.057** |
| | (0.024) |
| Week 2 Performance | 0.029* |
| | (0.015) |
| Week 3 Performance | 0.016 |
| | (0.016) |
| Week 4 Performance | −0.023* |
| | (0.013) |
| Treatment Arm: 2 | 0.050 |
| | (0.036) |
| Treatment Arm: 3 | −0.023 |
| | (0.036) |
| Treatment Arm: 4 | −0.022 |
| | (0.040) |
| Treatment Arm: 5 | −0.015 |
| | (0.036) |
| Treatment Arm: 6 | −0.033 |
| | (0.059) |
| Treatment Arm: 7 | −0.028 |
| | (0.048) |
| Treatment Arm: 8 | −0.027 |
| | (0.059) |
| Treatment Arm: 9 | −0.018 |
| | (0.039) |
| Treatment Arm: 10 | −0.016 |
| | (0.064) |
| Treatment Arm: 11 | −0.020 |
| | (0.059) |
| Treatment Arm: 12 | 0.207*** |
| | (0.048) |
| Treatment Arm: 13 | −0.011 |
| | (0.042) |
| Treatment Arm: 14 | −0.022 |
| | (0.054) |
| Treatment Arm: 15 | 0.147** |
| | (0.059) |
| Treatment Arm: 16 | −0.023 |
| | (0.044) |
| Constant | 0.027*** |
| | (0.033) |
| Observations | 647 |
| $R^2$ | 0.065 |
| Adjusted $R^2$ | 0.037 |
| Residual Std. Error | 0.143 (df = 627) |
| F Statistic | 3.313*** (df = 19; 627) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

Table 15

| | *Dependent variable:* |
|---|---|
| | dropped |
| Division: 1 | 0.028 |
| | (0.043) |
| Division 2 | −0.000 |
| | (0.048) |
| Division 3 | −0.000 |
| | (0.077) |
| Division 4 | 0.045 |
| | (0.042) |
| Division 5 | −0.000 |
| | (0.179) |
| Division 6 | −0.000 |
| | (0.130) |
| Division 7 | 0.125* |
| | (0.073) |
| Division 8 | 0.014 |
| | (0.045) |
| Division 9 | −0.000 |
| | (0.059) |
| Division 10 | 0.065 |
| | (0.043) |
| Division 11 | −0.000 |
| | (0.070) |
| Division 12 | −0.000 |
| | (0.108) |
| Division 13 | −0.000 |
| | (0.077) |
| Division 14 | 0.010 |
| | (0.043) |
| Division 15 | −0.000 |
| | (0.061) |
| Division 16 | −0.000 |
| | (0.082) |
| Division 17 | −0.000 |
| | (0.179) |
| Constant | 0.000 |
| | (0.040) |
| Observations | 737 |
| $R^2$ | 0.021 |
| Adjusted $R^2$ | −0.002 |
| Residual Std. Error | 0.174 (df = 719) |
| F Statistic | 0.895 (df = 17; 719) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

Table 16

| | *Dependent variable:* |
|---|---|
| | dropped |
| City 1 | 0.000 |
| | (0.146) |
| City 2 | 0.000 |
| | (0.174) |
| City 3 | 0.111 |
| | (0.130) |
| City 4 | 0.011 |
| | (0.125) |
| City 5 | 0.000 |
| | (0.146) |
| City 6 | 0.000 |
| | (0.140) |
| City 7 | 0.000 |
| | (0.151) |
| City 8 | 0.038 |
| | (0.124) |
| City 9 | 0.000 |
| | (0.151) |
| City 10 | 0.008 |
| | (0.124) |
| City 11 | 0.080 |
| | (0.128) |
| City 12 | 0.000 |
| | (0.174) |
| City 13 | 0.000 |
| | (0.214) |
| City 14 | 0.000 |
| | (0.214) |
| City 15 | 0.000 |
| | (0.214) |
| Constant | −0.000 |
| | (0.123) |
| Observations | 737 |
| $R^2$ | 0.015 |
| Adjusted $R^2$ | −0.006 |
| Residual Std. Error | 0.174 (df = 721) |
| F Statistic | 0.723 (df = 15; 721) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

# References


Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296.

Antonelli, J., Mealli, F., Beck, B., and Mattei, A. (2023). Principal stratification with continuous treatments and continuous post-treatment variables. *arXiv preprint arXiv:2309.14486*.

Aronow, P., Chang, H., and Lopatto, P. (2024). Randomization-based confidence intervals for the local average treatment effect. *arXiv preprint arXiv:2404.18786*.

Blattman, C., Jamison, J., Koroknay-Palicz, T., Rodrigues, K., and Sheridan, M. (2016). Measuring the measurement error: A method to qualitatively validate survey data. *Journal of Development Economics*, 120:99–112.

Bojinov, I., Chen, A., and Liu, M. (2020). The importance of being causal. *Harvard Data Science Review*, 2(3).

Bojinov, I., Choudhury, P., and N Lane, J. (2021). Virtual watercoolers: A field experiment on virtual synchronous interactions and performance of organizational newcomers. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (21-125).

Bojinov, I. and Shephard, N. (2019). Time series experiments and causal estimands: exact randomization tests and trading. *Journal of the American Statistical Association*, 114(528):1665–1682.

Bojinov, I., Simchi-Levi, D., and Zhao, J. (2022). Design and analysis of switchback experiments. *Management Science*.

Bojinov, I., Tu, Y., Liu, M., and Xu, Y. (2019). Causal inference from observational data: Estimating the effect of contributions on visitation frequency atlinkedin.

Branson, Z., Dasgupta, T., and Rubin, D. B. (2016). Improving covariate balance in 2 k factorial designs via rerandomization with an application to a new york city department of education high school study.

Buell, R. W., Kim, T., and Tsay, C.-J. (2017). Creating reciprocal value through operational transparency. *Management Science*, 63(6):1673–1695.

Byar, D. P., Herzberg, A. M., and Tan, W.-Y. (1993). Incomplete factorial designs for randomized clinical trials. *Statistics in Medicine*, 12(17):1629–1641.

Camuffo, A., Cordova, A., Gambardella, A., and Spina, C. (2020). A scientific approach to entrepreneurial decision making: Evidence from a randomized control trial. *Management Science*, 66(2):564–586.

Cappelli, P. and Conyon, M. J. (2018). What do performance appraisals do? *ILR Review*, 71(1):88–116.

Cloutier, O., Felusiak, L., Hill, C., and Pemberton-Jones, E. J. (2015). The importance of developing strategies for employee retention. *Journal of Leadership, Accountability & Ethics*, 12(2).

Cluff, L. A., Lang, J. E., Rineer, J. R., Jones-Jack, N. H., and Strazza, K. M. (2018). Training employers to implement health promotion programs: results from the cdc work@ health® program. *American Journal of Health Promotion*, 32(4):1062–1069.

Cox, D. R. (1958). Planning of experiments.

Dasgupta, T., Pillai, N. S., and Rubin, D. B. (2015). Causal inference from 2k factorial designs by using potential outcomes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(4):727–753.

Ding, P., Feller, A., and Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 655–671.

Dong, N. (2015). Using propensity score methods to approximate factorial experimental designs to analyze the relationship between two variables and an outcome. *American Journal of Evaluation*, 36(1):42–66.

Egami, N. and Imai, K. (2018). Causal interaction in factorial experiments: Application to conjoint analysis. *Journal of the American Statistical Association*.

Ellenbecker, C. H., Samia, L., Cushman, M. J., and Porell, F. W. (2007). Employer retention strategies and their effect on nurses' job satisfaction and intent to stay. *Home Health Care Services Quarterly*, 26(1):43–58.

Espinosa, V., Dasgupta, T., and Rubin, D. B. (2016). A bayesian perspective on the analysis of unreplicated factorial experiments using potential outcomes. *Technometrics*, 58(1):62–73.

Fisher, R. A. (1942). *The Design of Experiments*. New York: Hafner-Publishing, 3rd edition.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Galván Vela, E., Mercader, V., Arango Herrera, E., and Ruíz Corrales, M. (2022). Empowerment and support of senior management in promoting happiness at work. *Corporate Governance: The International Journal of Business in Society*, 22(3):536–545.

Gillen, B., Snowberg, E., and Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy*, 127(4):1826–1863.

Girotra, K., Terwiesch, C., and Ulrich, K. T. (2010). Idea generation and the quality of the best idea. *Management science*, 56(4):591–605.

Gubler, T., Larkin, I., and Pierce, L. (2018). Doing well by making well: The impact of corporate wellness programs on employee productivity. *Management Science*, 64(11):4967–4987.

Ham, D. W., Bojinov, I., Lindon, M., and Tingley, M. (2022). Design-based confidence sequences for anytime-valid causal inference. *arXiv preprint arXiv:2210.08639*.

Han, K., Basse, G., and Bojinov, I. (2024). Population interference in panel experiments. *Journal of Econometrics*, 238(1):105565.

Hart, D., Paetow, G., and Zarzar, R. (2019). Does implementation of a corporate wellness initiative improve burnout? *Western Journal of Emergency Medicine*, 20(1):138.

Hewitt, C. E., Kumaravel, B., Dumville, J. C., Torgerson, D. J., Group, T. A. S., et al. (2010). Assessing the impact of attrition in randomized controlled trials. *Journal of clinical epidemiology*, 63(11):1264–1270.

Jin, H. and Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, 103(481):101–111.

Kim, J., Candido, C., Thomas, L., and de Dear, R. (2016). Desk ownership in the workplace: The effect of non-territorial working on employee workplace satisfaction, perceived productivity and health. *Building and Environment*, 103:203–214.

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., and Pohlmann, N. (2013). Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1168–1176.

Kohavi, R., Tang, D., and Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press.

Koning, R., Hasan, S., and Chatterji, A. (2022). Experimentation and start-up performance: Evidence from a/b testing. *Management Science*, 68(9):6434–6453.

Levitt, S. D. and List, J. A. (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review*, 53(1):1–18.

Li, X., Ding, P., and Rubin, D. B. (2020). Rerandomization in 2^k factorial experiments.

Lu, J. (2016). Covariate adjustment in randomization-based causal inference for 2k factorial designs. *Statistics & Probability Letters*, 119:11–20.

Lu, J. and Deng, A. (2017). On randomization-based causal inference for matched-pair factorial designs. *Statistics & Probability Letters*, 125:99–103.

Mao, J. and Bojinov, I. (2021). Quantifying the value of iterative experimentation. *arXiv preprint arXiv:2111.02334*.

Marini, M. M., Olsen, A. R., and Rubin, D. B. (1980). Maximum-likelihood estimation in panel studies with missing data. *Sociological methodology*, 11:314–357.

Mattei, A. and Mealli, F. (2007). Application of the principal stratification approach to the faenza randomized experiment on breast self-examination. *Biometrics*, 63(2):437–446.

Maxwell, G., Rankine, L., Bell, S., and MacVicar, A. (2007). The incidence and impact of flexible working arrangements in smaller businesses. *Employee Relations*.

Mills, P. R., Kessler, R. C., Cooper, J., and Sullivan, S. (2007). Impact of a health promotion program on employee health risks and work productivity. *American Journal of Health Promotion*, 22(1):45–53.

Montgomery, D. C. (2017). *Design and analysis of experiments*. John Wiley & sons.

Mujtaba, B. G. and Cavico, F. J. (2013). Corporate wellness programs: implementation challenges in the modern american workplace. *International journal of health policy and management*, 1(3):193.

Mukerjee, R., Dasgupta, T., and Rubin, D. B. (2018). Using standard tools from finite population sampling to improve causal inference for complex experiments. *Journal of the American Statistical Association*, 113(522):868–881.

Ni Mhurchu, C., Aston, L. M., and Jebb, S. A. (2010). Effects of worksite health promotion interventions on employee diets: a systematic review. *BMC public health*, 10(1):1–7.

Pashley, N. E. and Bind, M.-A. C. (2023). Causal inference for multiple treatments using fractional factorial designs. *Canadian Journal of Statistics*, 51(2):444–468.

Pelet, J.-É. and Papadopoulou, P. (2012). The effect of colors of e-commerce websites on consumer mood, memorization and buying intention. *European Journal of Information Systems*, 21(4):438–467.

Radulescu, A. (2018). Users' social trust of sharing data with companies: online privacy protection behavior, customer perceived value, and continuous usage intention. *Contemp. Readings L. & Soc. Just.*, 10:137.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371):591–593.

Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: application to studies with" censoring" due to death. *Statistical Science*, pages 299–309.

Sagarin, B. J., West, S. G., Ratnikov, A., Homan, W. K., Ritchie, T. D., and Hansen, E. J. (2014). Treatment noncompliance in randomized experiments: statistical approaches and design issues. *Psychological methods*, 19(3):317.

Salminen, J., Kaate, I., Kamel, A. M. S., Jung, S.-g., and Jansen, B. J. (2021). How does personification impact ad performance and empathy? an experiment with online advertising. *International Journal of Human–Computer Interaction*, 37(2):141–155.

Schwartz, S. L., Li, F., and Mealli, F. (2011). A bayesian semiparametric approach to intermediate variables in causal inference. *Journal of the American Statistical Association*, 106(496):1331–1344.

Spence, G. B. (2015). Workplace wellbeing programs: if you build it they may not come... because it's not what they really need!

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472.

Sun, J., Zhang, D. J., Hu, H., and Van Mieghem, J. A. (2022). Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science*, 68(2):846–865.

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

Vicente-Herrero, T., Burke, T. A., and Laínez, M. J. (2004). The impact of a worksite migraine intervention program on work productivity, productivity costs, and non-workplace impairment among spanish postal service employees from an employer perspective. *Current medical research and opinion*, 20(11):1805–1814.

Wu, C. F. J. and Hamada, M. S. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. Wiley, New York, NY.

Yates, F. (1937). The design and analysis of factorial experiments.

Zhang, L., Rao, A., and Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

Zhao, A. and Ding, P. (2022). Reconciling design-based and model-based causal inferences for split-plot experiments. *The Annals of Statistics*, 50(2):1170–1192.

Zhao, A., Ding, P., Mukerjee, R., and Dasgupta, T. (2018). Randomization-based causal inference from split-plot designs.