ABSTRACT

WINPISINGER, RACHEL HANNAH. Improvements on Sequentially Weighted m-Factors-at-a-Time Experiments. (Under the direction of Jonathan Stallrich).

The emergence of internet-based services and products in the late 1990s created an unparalleled opportunity for online enterprises to implement extensive data-driven decision-making processes. This has led many industry leaders to devote significant resources to online controlled experiments (OCEs), sequential experiments used to evaluate the effects of modifications or new features of their products on both their clientele and overall business performance.

In OCEs and other applications, the task of synthesizing all potential treatment combinations can be quite labor-intensive, and certain treatments may lead to negative user experiences or financial losses. Multi-armed bandits (MABs), a popular choice for similarly designed experiments, do not handle these limitations well, leaving a need for alternative methods to address this problem. Larsen (2023) developed a sequential heuristic that focuses on exploration while successfully identifying the optimal treatment and avoiding harmful treatment combinations. It relies on an intuitive approach to parametrizing factorial effects, ensuring that unbiased estimations can be achieved with the replication of only a limited number of treatments, something impossible to achieve with traditional orthogonal effects without additional assumptions. In this thesis, we identify some drawbacks with this method and propose solutions to create a more robust heuristic. This new method is compared via simulation study with the original method and three MAB strategies that focus on minimizing regret. Our proposed method results in equal or lower regret and error rates as compared the other strategies examined.

Improvements on Sequentially Weighted m-Factors-at-a-Time Experiments

by
Rachel Hannah Winpisinger

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Statistics

Raleigh, North Carolina
2025

APPROVED BY:

Srijan Sengupta	Marie Davidian
William Rand	Jonathan Stallrich Chair of Advisory Committee

BIOGRAPHY

Rachel Hannah Winpisinger grew up in Greenwich, Connecticut. She attended The Pennsylvania State University, graduating in 2018 with a double major in Applied Statistics and Economics. After graduation, Rachel worked as a data scientist across many industries including retail marketing, mortgages and healthcare. In 2021, Rachel returned to school part-time as a master's student at North Carolina State Univeristy (NCSU). She successfully defended her thesis in March of 2025 and accepted an offer to continue her research in the doctoral statistics program at NCSU in the fall.

ACKNOWLEDGEMENTS

I would like to extend my deep thanks and gratitude to my advisor, Dr. Jonathan Stallrich, for his phenomenal support and guidance in writing this thesis. I could not ask for a better mentor and it has been an honor to work with him. I also would like to acknowledge my parents, Marcia and Kevin Winpisinger. Their patience and encouragement has been an unparalleled support. Special thanks goes to my teachers and mentors Dr. Sarah Goldin and Mr. Robert Brown. Your lessons have been instrumental in shaping my academic journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

IMPROVEMENTS ON SEQUENTIALLY WEIGHTED m-FACTORS-AT-A-TIME EXPERIMENTS

1.1 Introduction

Online controlled experiments (OCEs) are the digital adaptation of randomized controlled trials (RCTs), employing user-generated data to evaluate and enhance internet-based products and services (referred to as treatments throughout this paper). Predominantly, OCEs are real-time experiments that compare one or more test treatments against a control, or baseline, treatment, a methodology known in experimental design literature as treatment-vs-control experiments or TvCs. These experiments play a crucial role in driving financial success for numerous organizations and are increasingly utilized as standard tools in public health, governmental services, and political or social campaigns (Siroker and Koomen 2013). In short, OCEs are an incredibly influential tool for any company with a presence online, regardless of industry.

TvCs can quickly become cumbersome for factorial experiments that yield a relatively large number of potential test treatments. Suppose there are k factors each with $s_k \geq 2$ levels. Assuming no limitations in how the factor settings can be chosen, there are $\prod_k s_k$ possible treatment combinations. Even with $s_k = 2$, this number can grow for modest values of k and a large number of runs may be required to fully explore the treatment combinations. RCTs, however, are typified by a limited numbers of runs, large effect sizes and relatively more control over the factors of interest. OCEs boast a different structure and set of assumptions:

1. The cost to design and run an OCE is low, even negligible, but the effect sizes are relatively small. Companies have the ability to run larger scale-experiments to detect small effect sizes, which have large impact when aggregated over their large collection of users.

2. There are numerous examples where OCEs have less control over certain aspects of their process, or are hesitant to attempt certain treatment combinations. Consider an experiment concerning display ad copy for a new brand launch. The experimenter is

limited in their delivery of the various treatments by the search engine's ad placement algorithm. If certain copy is less prioritized by the algorithm, fewer people will be exposed to the treatment without any input from the experimenter.

3. OCEs are performed in a sequential fashion, with treatments assigned and applied to online users as they become relevant and engaged with the website.

4. OCE practitioners are concerned about replicating a large number of treatment combinations in case this results in a negative user experience. This concern is not only about the possibility of exposing users to poor treatments, but about having an overwhelming number of versions of a website that can confuse the user base.

These breaches from the typical position an experimenter finds themselves in indicate a need for a unique approach to handle OCEs. In particular, there is a need to balance optimization and learning. Optimization refers to the identification of the best treatment combination(s). Without loss of generality, we take this to mean identifying the treatment combination that yields the largest mean response. Conversely, learning encompasses the process of exploration aimed at uncovering significant factors or combinations of factors and understanding their influence on a specific system. Many industrial applications of OCEs require safe exploration among the treatments, particularly when practitioners aim to deliberately avoid certain combinations of settings. This need may arise for several reasons. One case is that the supporting of every conceivable factorial combination can entail substantial overhead and may prove impractical. Additionally, the arrival of sequential data in real-time necessitates that experimenters swiftly identify optimal treatments. Experiments altering more than a limited subset of settings may require excessive time for implementation (Optimizely 2022). Moreover, it is imperative to avoid treatment combinations that lead to negative outcomes, especially when they affect real individuals (such as users of a healthcare site) or may result in financial losses.

One popular framework to sequential optimization is the multi-armed bandit (MAB). MAB methods are well prepared to handle a large number of treatments and are usually designed to minimize the frequency of implementing suboptimal treatment combinations, thus favoring optimization over learning. The premise is that the experimenter sits at a slot machine with $k \geq 2$ arms, or treatments. Each time the experimenter pulls an arm, there is output providing both the immediate value of the output and information concerning that arm. The experimenter aims to balance the need to collect information on all the arms to identify which arm(s) maximize the output while minimizing the pulling of suboptimal arms. This process has two aspects, exploration of the entire factorial space to identify the optimal treatment and exploitation, or continuing replication, of the estimated optimal treatment. This framework fails to meet the goals of OCEs in two ways. The first is that due to the need to replicate all possible treatment combinations to identify the optimal choice, treatment combinations with negative outcomes will be explored. Further, these methods are guaranteed to require the development of all the treatment combinations, thus increasing the costs of the experiment. Additionally, the minimal exploration at the beginning of these methods does not provide much information regarding the treatments outside of the optimal. These methods specialize in identifying the best treatment and do not provide much information beyond it.

In addition to leaving possible information for the experimenter on the table, the MAB methods fail to utilize all available information collected in the algorithms. As a rule, they neglect to leverage common principles assumed with factorial experiments that lead to shared information between the many treatment combinations. Indeed, they assume that there is no information shared between the treatments at all, despite our knowledge that this is not the case. As such, these methods ignore the most effective route to successfully identifying their goal. Haizler and Steinberg (2021) addressed these issues using a sequential strategy that embraces factorial effect principles. They assumed all k factors took on 2 levels and the response was driven primarily by the main effects, assuming all interaction effects are negligible. This implies a simplified underlying structure of all possible 2^k treatment means (see Section 1.2.3), which allows information to be gained about all possible treatment means by only replicating some of them. They leverage the literature on orthogonal fractional factorial designs where the main effects can be efficiently estimated using only 2^{k-E} runs for some integer $1 \leq E < k$ such that $2^{k-E} > k+1$. Their strategy, referred to as FFD+TS, involves declaring E factors as active, and then using Thompson Sampling, a popular MAB method, to decide on the optimal settings of these factors. The remaining $k-E$ factors are explored using a 2^{k-E} fractional factorial design. Their sequential method performed well in simulation studies when the main effect model was correct, balancing exploration and exploitation better than Thompson Sampling alone.

One drawback of FFD+TS is it assumes all factor settings are easy to change so that any of the 2^k treatment combinations are feasible. For an OCE, it may be overwhelming to both the programmers and users to consider so many possible combinations. Moreover, another drawback of their method is they ignore the possibility of interactions in their model. Incorrectly specifying the model can lead to significant bias when estimating the main effects, which can lead to errors in identifying the set of factors to exploit and explore. They do not consider bias at any point in their paper. In response to these criticisms, Larsen (2023) proposed the Sequentially Weighted mFAT (SW mFAT) algorithm to bridge the gap between learning and optimization in OCEs. Similar to FFD+TS, the SW mFAT algorithm is split into stages that perform a factorial experiment on a subset of the treatment combinations. Unlike FFD+TS, each stage of the SW mFAT algorithm has a designated baseline treatment considered to be the current estimated optimum, and the goal is to perform inference on a subset of the so-called baseline factorial effects (see Section 1.2.6). A benefit of this new approach is the baseline effects are guaranteed to be estimated without bias using only those treatment combinations that change up to m factors from the baseline treatment, regardless of assumptions about high-order factorial effects. For each stage of SW mFAT, experimental units, arriving in sequential blocks, are allocated to the mFAT treatments to optimally learn about the relatively few baseline factorial effects of interest, given the current understanding of each effect's magnitude. After each stage, the baseline treatment is updated to what is estimated to be the best treatment combination and a new stage is performed.

In contrast to the MAB methods, each stage of SW mFAT focuses on learning to safely identify candidates that improve over the baseline. As such, once an effect is found to be of sufficient magnitude, there is no value in further learning about the effect. This often leads to halting replication of some treatment combinations. Fortunately, the mean of the baseline treatment is a core part of each baseline effect, which causes it to be the most replicated treatment combination. This has the appealing practical property of exposing most users to a safe, reliable treatment, and preventing exposing users to too many treatment combinations in a stage. To minimize risk when updating the baseline treatment, one could choose the best treatment combination to be the one with the largest observed sample mean. On the other hand, baseline factorial effects can use the information from these mFAT experiments to estimate means of unobserved treatments, although these estimates are only unbiased under certain assumptions of higher order interactions.

While successful in balancing the opposing desires of learning and optimization while working successfully in the limitations of an OCE, there are some pitfalls with SW mFAT which we examine in Section 1.3. These can be split in two categories: the suboptimal allocation of experimental units to treatments and a propensity for error in certain situations. A solution to these pitfalls is detailed in Section 1.4, in the form of a new proposed heuristic, Improved SW mFAT, that incorporates guide rails to minimize unmitigated errors.

The paper is structured as follows. Section 1.2 introduces notation, briefly reviews multi-armed bandits and optimal design theory, baseline parametrization, and the premise of

mFAT designs along with methods for choosing optimal mFAT designs under baseline parametrization. Section 1.3 discusses the SW mFAT algorithm proposed by Larsen (2023) and the drawbacks to this method. Section 1.4 proposes a new heuristic, Improved SW mFAT, which addresses the drawbacks identified in the previous section. The learning and optimization properties of these heuristics are explored via simulation in Section 1.5 and compared to popular MAB algorithms, with future work suggested in Section 1.6.

1.2 Background

Suppose there are k two-level experimental factors, $F_1, ..., F_k$, each with the potential to impact the response Y_i of $i = 1, ..., N$ experimental units. When F_l is set to a baseline representing a no change or control setting let $j_l = 0$. Conversely, let $j_l = 1$ when F_l is set to a specific intervention setting. A combination of factors (treatments) is represented by the k-tuple $j_1 j_2 ... j_k$, with 2^k treatments in total. Treatment means are denoted as $\mu_{j_1 j_2 \cdots j_k}$. The response is modeled as follows:

$$Y_i = \mu_{j_1^i j_2^i \cdots j_k^i} + \epsilon_i \tag{1.1}$$

where $\mu_{j_1^i j_2^i \cdots j_k^i}$ is the expected value of Y_i and $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. For example, if $k = 3$, then a single arbitrary run that received treatments F_1 and F_3 has an expected response of $\mathbb{E}[Y_i] = \mu_{101}$.

1.2.1 Sequential Design: Multi Armed Bandits

Define the combination of factors that yields the largest mean response as the treatment combination $j_1^* j_2^* \ldots j_k^*$ so $\mu_{j_1^* j_2^* \cdots j_k^*} \geq \mu_{j_1 j_2 \cdots j_k}$ for all 2^k treatments. Combined with the sequential structure of OCEs, the problem of identifying $j_1^* j_2^* \ldots j_k^*$ is very similar to the MAB framework (Lai et al. 1985). The hypothetical "bandit" is a slot machine with $p = 2^k$ arms (treatments). At each step, the gambler is tasked with selecting the arm which would maximize their winnings, or more formally, with choosing the treatment that has the largest increase in mean over the baseline.

The situation posed by the MAB is one of the clearest demonstrations in literature of the balance between optimization and learning. As the gambler pulls more arms, more information is obtained, but at the price of potentially pulling an arm that hurts the final outcome. The heuristics discussed later in this paper are designed to balance this trade-off by using sequential, factorial analysis frameworks, whereas most MAB solutions focus more heavily on optimizing regret (Scott 2010; Haizler and Steinberg 2021; Audibert et al. 2009; Auer et al. 2002; Mak et al. 2022), the observed loss in reward after analyzing t experimental units:

$$R_t = \frac{1}{t}\sum_{n=1}^{t}(\mu_{j_1^* j_2^* \ldots j_k^*} - Y_n). \tag{1.2}$$

In Section 1.5 the heuristic proposed by Larsen (2023) and the improved version of it proposed in this paper are compared to three well-known MAB algorithms with respect to regret: Thompson Sampling (TS), Upper Confidence Bound for Normal Distributions (UCB-Normal), and Variance-Aware Upper Confidence Bound (UCB-V).

The TS algorithm is a Bayesian approach introduced by Thompson (1935) and later popularized in Scott (2010, 2015). It assumes an uninformed set of priors on the reward for each treatment, a (which are assumed to be Normal(μ_a,σ_a^2). As both the mean and variance of the reward is unknown, the conjugate prior is the normal-inverse-gamma distribution such that

$$P(\mu_a, \sigma_a^2 | m_a, \nu_a, \alpha_a, \beta_a) \sim N - \Gamma^{-1}(m_a, \nu_a, \alpha_a, \beta_a) \tag{1.3}$$

where m_a, ν_a, α_a and β_a are the hyperparameters and represent the prior mean (μ_a), the prior "count" ($\nu_a > 0$), prior shape parameter ($\alpha_a > 0$) and prior scale parameter ($\beta_a > 0$).

At each time step, t, in the algorithm an action, A_t, is chosen and a reward, X_t, is observed. For purposes of computation let $D_{t,a} = \{X_\tau : A_\tau = a, \tau = 1, ..., t-1\}$ and $N_t(a) = |\mathscr{D}_{t,a}|$. Since we assume that the arms are independent, we can calculate the posterior for each arm, a, for the set of rewards received up to but not including time t as

$$P(\mu_a, \sigma_a^2 | \mathscr{D}_{t,a}) \propto \mathscr{N}(\mu_a; \rho_{t,a}, \zeta_a^2)\, \mathscr{IG}(\sigma_a^2; \tfrac{1}{2}N_t(a) + \alpha_a, \beta_{t,a}) \tag{1.4}$$

where

$$\begin{aligned}
\rho_{t,a} &= \frac{\nu_a m_a + N_t(a)\bar{x}_{t,a}}{\nu_a + N_t(a)} \\
\zeta_a^2 &= \frac{\sigma_a^2}{N_t(a) + \nu_a} \\
\beta_{t,a} &= \beta_a + \tfrac{1}{2}s_{t,a} + \frac{N_t(a)\nu_a(\bar{x}_{t,a} - m_a)^2}{2(N_t(a) + \nu_a)}.
\end{aligned}$$

When an experimental unit arrives, the algorithm proceeds as follows:

1. For each arm of $a \in \mathbb{A}$

    - Draw $\hat{\sigma}_a^2 \sim \mathscr{IG}(\frac{1}{2}N_t(a) + \alpha_a, \beta_{t,a})$.
    - Compute $\hat{\zeta}_a^2 = \hat{\sigma}_a^2 / (N_t(a) + \nu_a)$.
    - Draw $\hat{\mu}_a \sim \mathscr{N}(\rho_{t,a}, \hat{\zeta}_a^2)$.

2. Select the arm with the highest expected reward:

$$A_t = \underset{a \in \mathscr{A}}{\arg\max}\, \mathbb{E}[X_t | \hat{\mu}_a, \hat{\sigma}_a^2] = \underset{a \in \mathscr{A}}{\arg\max}\, \hat{\mu}_a. \tag{1.5}$$

3. Assign the experimental unit (pull the arm) to A_t and observe reward X_t.

4. Update $N_t(a)$, $\bar{x}_{t,a}$, $\rho_{t,a}$, $s_{t,a}$, and $\beta_{t,a}$ for $a = A_t$.

This process is repeated for all experimental units. Though simple, TS can achieve optimal logarithmic expected regret and has been used in online applications (Agrawal and Goyal 2012; Amadio 2020).

The UCB-Normal (Auer et al. 2002) and UCB-V (Audibert et al. 2009) algorithms are part of a family of frequentist approaches that define upper confidence bounds on the expected response from each treatment in an effort to minimize regret. Experimental units are assigned to the treatment with the maximum upper confidence bound. These bounds are recomputed as more information about each treatment is obtained in such a manner as to focus optimization efforts on the best performing treatments but still allow uncertainty to permit sufficient exploration of all treatments.

Per its name, the UCB-Normal algorithm utilizes an upper confidence bound that assumes normally distributed responses from treatments with unknown mean and variance. The algorithm begins with a burn-in learning period during which each machine needs to be played a set number of times. This period is purely about exploration and no steps are taken to lower regret. After collecting this information, the algorithm switches to its primary process that employs a formula derived from the index-based policy of Agrawal (1995). The index of this policy is the sum of two terms. The first is the current average reward. The second relates to the size (according to Chernoff-Hoeffding bounds) of the one-sided confidence interval for the average reward within which the true expected reward falls with overwhelming probability.

The algorithm goes as follows for each experimental unit $i = 1, 2, \ldots$ with step two displayed graphically in Figure 1.1:

1. If there is a machine which has been played fewer than $8 \log(n)$ times then play this machine.

2. Otherwise play machine j that maximizes

$$\bar{y}_j + \sqrt{16\log(t-1) \times \frac{s_j^2}{t_j}} \tag{1.6}$$

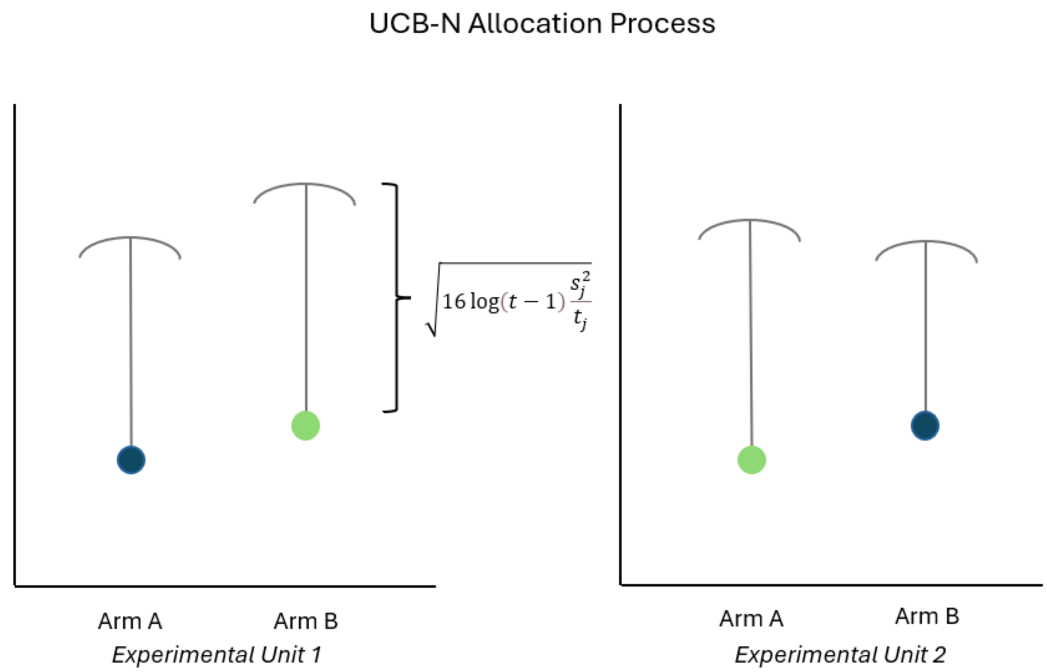
   where t indicates the current number of units in the experiment, t_j is the current number of units assigned to treatment j, $\bar{y}_j$ the sample mean of the units' response under treatment j, and $s_j^2 = \frac{\sum_{i=1}^{t_j}(y_{ij}-\bar{y}_j)^2}{t_j-1}$ is the sample variance of treatment i.

3. Update $\bar{y}_j$ and s_j^2 with the observed reward x_i

The one sided confidence interval is calculated using two components. The first is the sample variance. The second is an ever, albeit slowly, increasing critical value $\sqrt{16\log(t-1)}$. Since this value relies on t, it prevents the confidence interval from ever approaching 0 as the sample size increases.

The process encourages some exploration due to the large nature of the confidence bounds. As new information is acquired for one treatment, the estimated variance decreases which in turn lowers the bound for that treatment. This gives the opportunity for a treatment with a similar, but slightly lower sample mean to be selected for the next allocation in the process. Figure 1.1 demonstrates this process.

Figure 1.1: UCB-N allocation decision. After an experimental unit is assigned to Arm B, the associated variance decreases which allows Arm A to produce a higher confidence bound and be subsequently explored. The treatment with the green marker is assigned the experimental unit in each step.



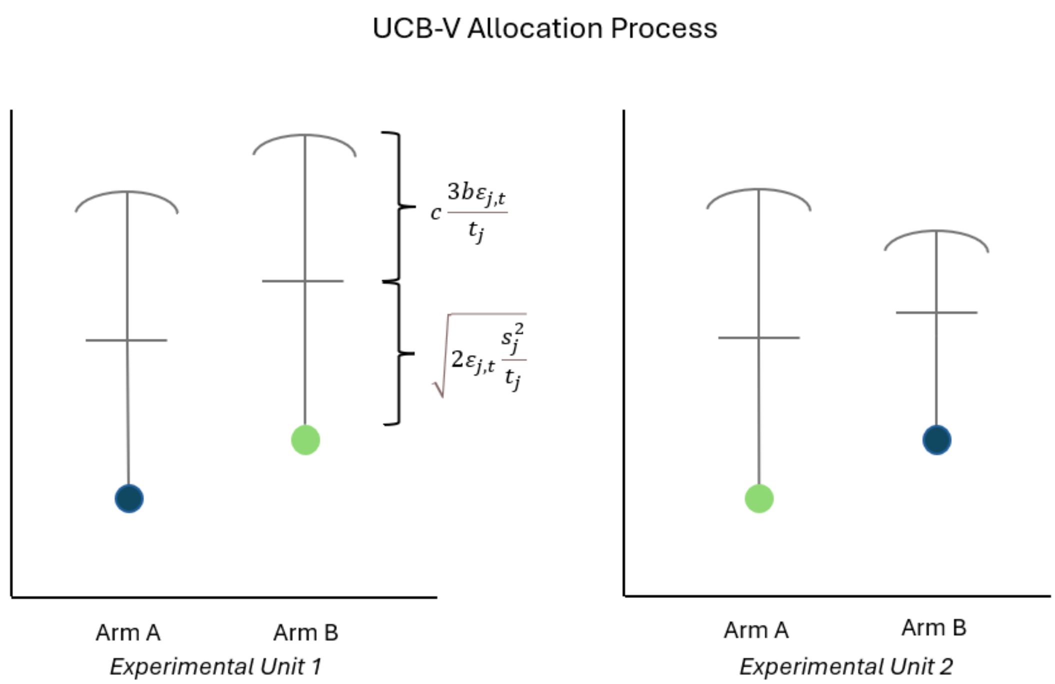
As noted in Audibert et al. (2009), the UCB-Normal algorithm achieves very desirable results assuming one knows*a priori* the distribution of the response is Normal. In many situations this cannot be determined, thus Audibert et al. (2009) suggest UCB-V, a more flexible bound that incorporates estimates of the response variance,

$$\bar{y}_j + \sqrt{2\mathscr{E}_{j,t}\frac{s_j^2}{t_j}} + c\,\frac{3b\mathscr{E}_{j,t}}{t_j} \tag{1.7}$$

where s_j^2 is the sample variance of the previous responses under treatment j, b is an upper bound on the response, and $\mathscr{E}_{j,t}$ and c are parameters that can be altered to control the algorithm's exploration behavior. This algorithm is equivalent to UCB-N if the third term is

0 and $2\mathscr{E}_{j,t} = \sqrt{16\log(t-1)}$, the critical value. The third term introduces a tunable value which decreases as treatment j is replicated. This value, whose influence is determined by c, is added to $\bar{y}_j$ to encourage more exploration. It augments the implicit exploration in UCB-N of treatments with lower sample means. It also encourages exploration of treatments with a lower variance, which do not lend themselves to having large confidence bounds. Figure 1.2 displays this phenomenon with the third term added.

Figure 1.2: UCB-V allocation decision. Similar to the UCB-N, exploration occurs by decreasing the confidence bound arms as they are exploited. By adding the third term, the amount of exploration can be more accurately controlled. The treatment with the green marker is assigned the experimental unit in each step.

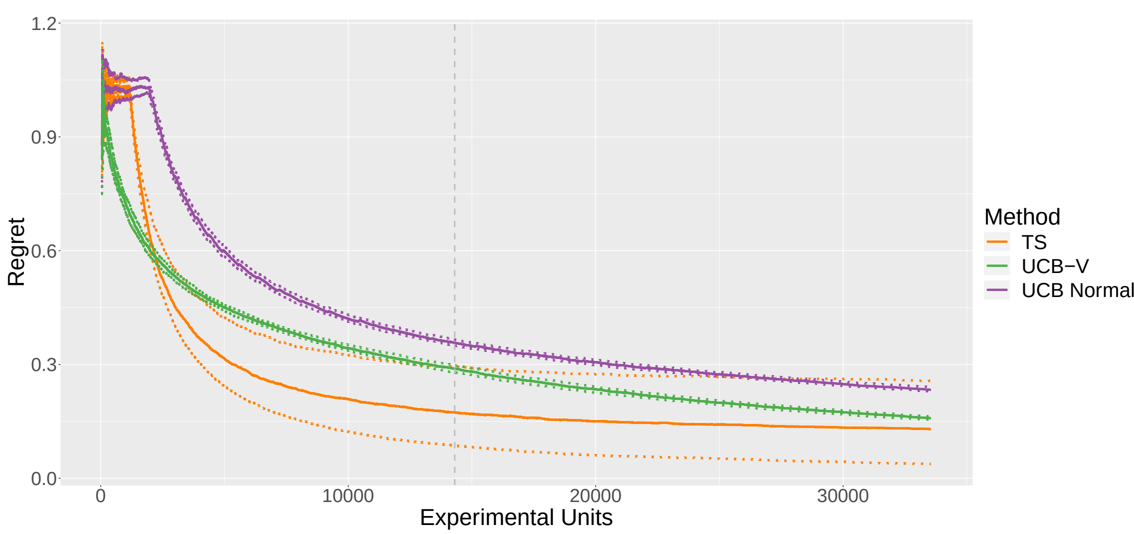


Audibert et al. (2009) selected $b = 10$ such that the probability of the normally distributed responses exceeding b is negligible, with $\mathscr{E}_{j,t} = 1$ and $c = 1$ in their proposal. Both UCB algorithms provide increased exploration which causes them to converge slower than TS to a minimized regret. However, due to the increased exploration, these methods have less variability in their regret functions.

There are two key drawbacks to these algorithms. First, they do not prioritize safe exploration of the factor space, particularly in the initial stages of the experiments, since they tend to over replicate the best observed treatment. TS exhibits this more extremely than the other two methods and the nuances of that are examined below. Second, they assume there is no shared information between the treatment means.

The first consequence of not prioritizing safe exploration is the susceptibility to mishandling treatments with extreme values early in exploration. This may result from large variances as compared to the effect size in question or simply random chance. Take for example an experiment with $k = 3$ in which the treatment means are $\mu_{100} = 3.2$,$\mu_{010} = 3.1$, $\mu_{000} = \mu_{001} = 3$ and all remaining treatment means equal $\mu_{000} = 3$. We performed a simulation study that performed each MAB method 100 times under these settings. Figure 1.3 shows the 25^{th}, 50^{th} and 75^{th} quantiles of the regret functions for the three MAB methods.

Figure 1.3: 25^{th}, 50^{th} and 75^{th} quantiles of the regret functions for Thompson Sampling, UCB-V and UCB-N. Thompson Sampling's median regret is the lowest but at the cost of a higher variability.



The three MAB methods all spend the beginning of the experiment in a burn in phase. While the two UCB methods explicitly call for this exploration period in their algorithms, the same result occurs naturally in TS. Since all the priors are assumed to be equal, each treatment is selected roughly equally until enough pulls are made to redefine the distributions to start favoring the treatment which minimizes regret. However, not all methods stay in this burn in phase for the same amount of time. Consider a hypothetical scenario in which after this initial burn in period, the four treatment mean estimates were $\hat{\mu}_{000} = 2.98, \hat{\mu}_{100} = 2.8, \hat{\mu}_{010} = 3.12,$ and $\hat{\mu}_{001} = 3.03$. All three MAB algorithms will weight their allocation towards 010 as the optimal treatment. Over time, the estimate of 010 will decrease to approach its expected value, $\mu_{010} = 3.1$. This increases the chance for another treatment with a similar estimated mean to receive an allocation. However, it is less likely for treatment 100 to be replicated, as its estimated mean is too low in comparison. Thus treatment 100 is never given the opportunity to recover to its expected mean. The UCB-N and UCB-V methods take steps to address this by having an explicit and equal burn in period for each treatment at the beginning of the experiment and allow for further course correction via the introduction of the confidence intervals. However, the Thompson Sampling method is very susceptible. The repercussions of this can be seen in the high variability in regret that Thompson Sampling has as compared to the other two methods depicted in Figure 1.3. While Thompson Sampling's median regret does come in lower than the two UCB methods, the 25^{th} and 75^{th} percentile lines show the disparity in possible outcomes. By the end of the experiment, slightly more than 25% of the Thompson Sampling procedures returned a higher regret than UCB-N and just under 50% resulted in a higher regret than UCB-V. As many of these experiments are only completed once, this variability in regret and thus in identifying the optimal result can have a hugely adverse impact on the experimenter.

The MAB methods can also susceptible to another problematic outcome: incorrectly identifying a treatment as having a significantly positive effect size when it does not. Consider another $k = 3$ experiment but with treatment means all equal to 3. Let the estimated treatment means after initial exploration be $\hat{\mu}_{000} = 2.98, \hat{\mu}_{100} = 3.5, \hat{\mu}_{010} = 3.02,$ and $\hat{\mu}_{001} = 2.99$. In this case, the models will target 100, having ascertained that this is the largest positive effect size. Eventually, all three models will collect enough data on 100 to revise the estimate down to somewhere closer to 3, much more in line with reality and the other treatments. However, it is possible that the extreme value is enough to 'pull' the estimate so that it remains measurably higher than all the other treatment combinations being examined. This can result in the model confirming that the optimal treatment is 100, when in actuality there is not an optimal treatment. For businesses, this can have the costly impact of overhauling a marketing campaign when in fact they could have used what was already in place for equal results.

Another repercussion of emphasizing exploitation over exploration is the loss of information. As mentioned above, OCEs tend to identify small and incremental improvements from a large array of simultaneously run treatments. By forcing the selection of and putting a focus on one top performer, the MAB methods lose the information about other treatments, even if they demonstrate an improvement over the control. As practitioners within the OCE community have noted, "failed experiments" or those that do not improve over the baseline, may still provide value with respect to future experimentation (Thomke 2020). Since the MAB algorithms explore just enough to ensure they have found the optimal treatment, information on non-optimal treatments is not prioritized and sometimes barely collected.

1.2.2 Factorial Analysis

The three MAB methods we have discussed all fail to leverage similarities between treatment combinations arising from their factorial nature. This forces the methods to examine each treatment combination in isolation with no information shared to find efficiencies in the process. This is a very conservative assumption, especially in the OCE environment. Embracing the factorial structure creates opportunities for more interconnected and efficient ways of estimating treatment means. Two such frameworks we will discuss are the orthogonal and baseline parametrization of factorial effects, which are both linear transformations of the $\mu_{j_1,\dots,j_k}$.

We begin with the orthogonal parameterization (OP), as it is the traditional definition of factorial effects. There are 2^k such effects and so they may also be indexed by $j_1,\dots,j_k$. The OP intercept, denoted $\theta^{OP}_{00\dots0} = 2^{-k}\sum \mu_{j_1,\dots,j_k}$ is the average of all 2^k means. The remaining effects are defined by $k-1$ sets of mutually orthogonal contrasts that involve all 2^k means. The OP main effect of F_1 is

$$\theta^{OP}_{10\dots0} = \frac{1}{2^{k-1}}\sum_{j_2=0}^{1}\cdots\sum_{j_k=0}^{1}(\mu_{1j_2\dots j_k} - \mu_{0j_2\dots j_k}) = \bar{\mu}_{1\dots} - \bar{\mu}_{0\dots}\ ,$$

being the expected difference in responses when F_1 is set to $j_1 = 1$ versus $j_1 = 0$, after averaging over the levels of all $k-1$ other factors. The main effects for the other factors are defined similarly, with averaging always taken over the other factors. The intention of a main effect is to estimate the marginal effect of a factor. However, such quantities can misrepresent the importance of a factor if its effect changes depending on the levels of another factor. We measure such inconsistent effects through so-called interaction effects. Interaction effects are defined as contrasts for groups of $2,3,\dots,k$ factors. For example, the two-factor interaction between F_1 and F_2 is defined by

$$\begin{aligned}\theta^{OP}_{110\dots0} &= \frac{1}{2^{k-2}}\sum_{j_3=0}^{1}\cdots\sum_{j_k=0}^{1}(\mu_{11j_3\dots j_k} - \mu_{01j_3\dots j_k}) - (\mu_{10j_3\dots j_k} - \mu_{00j_3\dots j_k}) \\ &= (\bar{\mu}_{11\dots} - \bar{\mu}_{01\dots}) - (\bar{\mu}_{10\dots} - \bar{\mu}_{00\dots})\ .\end{aligned}$$

The first term for $\theta^{OP}_{110\dots0}$, $(\bar{\mu}_{11\dots} - \bar{\mu}_{01\dots})$, is the expected difference in responses when F_1 is set to $j_1 = 1$ versus $j_1 = 0$, but with $j_2 = 1$ and averaging over the levels of all $k-2$ other factors. The second term is similar, but with $j_2 = 0$. If the effect of F_1 were consistent across the levels of F_2, then $\theta^{OP}_{110...0} = 0$. Three-factor and higher order interactions are defined similarly, although with cumbersome expressions. The key idea for an OP m-factor interaction effect is that it measures the consistency of a given factor's effect on the response across the $m-1$ other factors, and then averages over the remaining $k-m$ factors.

The baseline parameterization (BP) of factorial effects (Mukerjee and Tang 2012) is defined nearly identically to the OP effects except rather than averaging over the levels of the uninvolved factors, the levels of the factors are fixed to their baseline setting $j_l = 0$. Hence the BP intercept is simply $\theta^{BP}_{00...0} = \mu_{00...0}$, the baseline main effect for F_1 is $\theta^{BP}_{10...0} = \mu_{10...0} - \mu_{00...0}$, and the BP two-factor interaction between F_1 and F_2 is

$$\theta^{BP}_{110\cdots0} = (\mu_{110\cdots0} - \mu_{010\cdots0}) - (\mu_{100\cdots0} - \mu_{000\cdots0}) \,.$$

Therefore, baseline factorial effects can be seen as an appropriate generalization of TvC experiments when the treatments arise from a factorial system.

Both factorial effect systems are defined as linear transformations of model 1.1. The vectorized version of (1.1) is

$$y = X\mu + \epsilon \tag{1.8}$$

where X is an $n \times 2^k$ matrix. The columns of X and the elements of μ are indexed by $j_1 \dots j_k$. The i-th row of X consists of a single 1 corresponding to the $j_1 \dots j_k$-th element and zeros elsewhere. After proper sorting of the elements of μ, and ignoring the $1/2^{k-m}$ averaging constants, the vector of the OP factorial effects are by $\theta_{OP} = H_{OP}\mu$ where

$$H_{OP} = Z_1^{OP} \otimes \cdots \otimes Z_k^{OP} \qquad \text{and} \qquad Z_l^{OP} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} . \tag{1.9}$$

The name OP comes from the fact that $H_{OP}H_{OP}^T = I$ and so the θ^{OP} decompose μ into orthogonal components. Per Stallings and Morgan (2015), a similar transformation matrix H_{BP} exists for the BP factorial effects, $\theta_{BP} = H_{BP}\mu$, using

$$H_{BP} = Z_1^{BP} \otimes \cdots \otimes Z_k^{BP} \qquad \text{and} \qquad Z_l^{BP} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}. \tag{1.10}$$

Notably H_{BP} is not an orthogonal matrix. However, it is a lower-triangular matrix, as it is a Kronecker product of k lower triangular matrices. This means that many of the BP effects involve a small subset of all possible means. This is quite different from H_{OP}, where every element is ± 1.

Clearly, to estimate all 2^k OP or BP factorial effects, one needs to replicate all 2^k treatment combinations at least once. For such an experiment, the least-squares estimator for θ_{OP} is $\hat{\theta}_{OP} = H_{OP}\hat{\mu}$, where $\hat{\mu}$ is the observed means of all 2^k treatments, with variance $\text{Var}(\hat{\theta}_{OP}) \propto H_{OP}R^{-1}H_{OP}^T$ where R is a diagonal matrix of the replications of the 2^k treatments. Similar expressions exist for the BP effects, but with H_{OP} replaced by H_{BP}. If one were to run such an experiment, they would want to choose R to minimize the variances. In the case of θ_{OP}, this is uniquely accomplished by an equally-replicated experiment. For θ_{BP}, there is no clear winner as to what the best design would be. Stallings and Morgan (2015) construct designs that minimize $\text{tr}(H_{BP}R^{-1}H_{BP}^T)$ and so minimize the average variance of the 2^k BP factorial effects. Such designs do not equally replicate the treatment combinations, and tend to focus on replicating the baseline treatment and the treatments manipulating fewer factors at a time.

1.2.3 Fractional Factorial Designs

Experiments replicating all 2^k treatment combinations are quite large, especially considering the OCE assumption of small effect sizes. Fortunately, there are common principles assumed that make the design and analysis more practical (Wu and Hamada 2011):

1. **Effect hierarchy**: Most of the variation of the response can be explained by main effects and lower order interactions (usually two-factor interactions). The magnitude of main effects tend to exceed those of interactions.

2. **Effect heredity**: For an m-factor interaction effect to be important, so must be the main effect of at least one of its m involved factors. Strong heredity says that all m factors must have important main effects.

3. **Effect sparsity**: Most of the response's variation can be explained by relatively few effects.

4. **Factor sparsity**: When many factors are being studied, say $k \geq 4$, many of the factors and their corresponding factorial effects will not be important in explaining the response's variation.

These principles are empirically-motivated, primarily from the large number of agricultural and manufacturing experiments that have long-embraced factorial treatments. The principles present opportunities to simplify both the experiment and the analysis. In particular, they suggest that one may assume *a priori* that such three-factor or higher-order interaction effects are essentially 0 and need not be estimated at all. Therefore, fewer treatment combinations need to be replicated to estimate the lower-order effects of interest.

A fractional factorial design is one in which not all 2^k treatment combinations are replicated. The experiments are cost-effective, but can present significant challenges with unbiased estimation of the factorial effects of interest. In the case of θ_{OP}, let θ_A be the set we think could be nonzero and hence want to estimate. Let θ_I be the remaining effects which we presume equal 0. Then we can rewrite model (1.8) as

$$y = XH_{OP}^{-1}H_{OP}\mu + \epsilon = X\theta_{OP} + \epsilon = X_A\theta_A + X_I\theta_I + \epsilon \ . \tag{1.11}$$

If we assume $\theta_I = 0$, then the least-squares estimator for θ_A is $\hat{\theta}_A = (X_A^T X_A)^{-1} X_A^T y$ with variance $\mathrm{Var}(\hat{\theta}_A) \propto (X_A^T X_A)^{-1}$. We could then choose a design that minimizes $(X_A^T X_A)^{-1}$ in some sense to improve our estimation of θ_A. The same approach also holds for BP factorial effects, with H_{OP} replaced by H_{BP}.

The $\theta_I = 0$ assumption implicitly says that there is some shared information across the 2^k treatment means, and that all 2^k means can be recovered through another linear transformation $\mu = M\theta_A$. This is possible for OP factorial effects, but it turns out to be much more straightforward for BP factorial effects. Note that by definition $\mu = H_{BP}^{-1}\theta_{BP}$ where $\theta_{BP} = H_{BP}\mu$, using

$$H_{BP}^{-1} = Z_1^{BP,-1} \otimes \cdots \otimes Z_k^{BP,-1} \qquad \text{and} \qquad Z_l^{BP,-1} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} . \tag{1.12}$$

Therefore, any treatment mean can be written as the sum of a subset of all the baseline effects. Suppose we want to write this expression for some $j_1 \dots j_k$. Mukerjee and Tang (2012) and Stallings and Morgan (2015) show that the collection of baseline effects, $\theta_{j_1 \dots j_k}$, involved in the sum include the baseline effect with the same index $j_1 \dots j_k$ and all those that replace at least one $j_l = 1$ in $j_1 \dots j_k$ with $j_l = 0$. For example, for the 2^3 experiment

$$\mu_{110} = \theta_{110} + \theta_{100} + \theta_{010} + \theta_{000} \ . \tag{1.13}$$

If some $\theta_I = 0$, they would just simply need to be removed from this expression. For example, if in the 2^3 experiment all BP interaction effects were not important then $\theta_{110} = 0$ and

$$\mu_{110} = 0 + \theta_{100} + \theta_{010} + \theta_{000} = \theta_{100} + \theta_{010} + \theta_{000} \ . \tag{1.14}$$

Therefore, $\hat{\mu}_{110} = \hat{\theta}_{100} + \hat{\theta}_{010} + \hat{\theta}_{000}$ is unbiased for μ_{100}. We refer to this as an extrapolating estimator for μ_{110}.

The unbiasedness of $\hat{\theta}_A$ and any extrapolating estimator for a treatment mean relied on the assumption $\theta_I = 0$. If this does not hold, then

$$\mathbb{E}(\hat{\theta}_A) = \theta_A + (X_A^T X_A)^{-1} X_A^T X_I \theta_I \ . \tag{1.15}$$

If $(X_A^T X_A)^{-1} X_A^T X_I \theta_I \neq 0$, then any gains we get in minimizing variances would be useless as our estimates would be biased, and these biases would propagate to the extrapolating estimators. However, we would be protected against any $\theta_I \neq 0$ if the so-called aliasing matrix $(X_A^T X_A)^{-1} X_A^T X_I$ was shown to be 0. It is desirable then to minimize both variances and the magnitude of the aliasing matrix when choosing a fractional factorial design.

Much of the existing literature on fractional factorial designs that try to minimize variance and bias has focused on θ_{OP} and θ_A comprised of the main effects. Wu and Hamada (2011) and the references therein cover many of the popular techniques including regular fractional factorial designs and minimum aberration designs. Such techniques were employed in Haizler and Steinberg (2021) for the factors deemed unimportant. These ideas have been recently extended to BP effects.

Minimizing bias for θ_{OP} is complicated because the definition of each effect involves all 2^k treatment means. This is not the case of θ_{BP}. Indeed, each BP main effect involves only $\mu_{00\ldots0}$ and the mean of some treatment that manipulates only one factor from its baseline. Therefore, we are guaranteed to estimate all BP main effects without bias with a design that replicates the baseline treatment and the k treatments that set only one factor to $j_l = 1$ (i.e., where $\sum_l j_l = 1$). This is called a one-factor-at-a-time (OFAT, or 1FAT) design since we change at most one factor from its baseline setting. However, this ignores our ability to minimize variances. Research into BP effects has developed significantly in the last 15 years, with attention given to fractional factorial designs that minimize some combination of variance and bias.

Mukerjee and Tang (2012) proposed a minimum K-aberration criterion for selecting optimal two-level fractional designs using BP effects. The authors examine a design space limited to orthogonal arrays of strength 2, guaranteeing universally minimized variance estimates for the main effects within a main effects-only framework. They then quantify the bias introduced to the main effect estimates by the presence of interaction terms. The optimal K-aberration design was defined as one that sequentially minimized bias, starting from the lower-order interaction. Said optimal design was identified via a complete search algorithm, enhanced by the efficiency improvements proposed by Li et al. (2014). Mukerjee and Huda (2016) build upon Mukerjee and Tang (2012) by offering approximate theoretical frameworks for general factorial designs with arbitrary levels and run sizes. Meanwhile Mukerjee and Tang (2016) extended their own work by focusing on minimum aberration designs applicable to scenarios with large k values. Miller and Tang (2016) discuss heuristics for creating baseline designs that exhibit favorable K-aberration properties using typical 2^{k-p} designs. The main effect model presented by Mukerjee and Tang (2012) is further developed to incorporate situations with some important two-factor interaction effects in Chen et al. (2021). This research has been expanded to encompass three-level (Yan and Zhao 2023) and s-level minimum aberration baseline designs (Yan and Zhao 2025).

Karunanayaka and Tang (2017); Li et al. (2022) explored lowering the bias of estimates by finding a middle ground between minimal aberration design and OFAT designs. As mentioned earlier, OFAT designs guarantee unbiased BP main effect estimates, but with a trade-off in variance. By combining these two types of designs, the authors were able to reduce the variance of OFAT estimates, while producing a nonzero but lower bias than that found by Mukerjee and Tang (2012). These designs were limited to baseline main effects and, to successfully lower the variance incurred below that of OFAT of Mukerjee and Tang (2012)'s designs, required in depth prior knowledge concerning the distributions of both σ^2 and effect sizes.

Despite comments found in Haizler and Steinberg (2021) and Kohavi et al. (2020), we are hesitant to accept the assumption that interactions are not important in OCEs. For one, this assumption and the other effect principles have mainly come from controlled experiments in agriculture and manufacturing. OCEs are inherently social experiments that are less understood. Moreover, the effect sizes for OCEs are already small, and even if the main effects are relatively large, an aggregate of many smaller higher order interactions could be enough to significantly bias estimates of the main effects. This leads us to immediately reject the OP factorial analysis for OCEs, since their bias reduction properties require the assumption of $\theta_I = 0$. Moreover, to reduce risk with estimating lower order BP factorial effects with bias, we agree with Larsen (2023)'s proposal to focus on m-factors-at-a-time designs, which is an extension of the OFAT design.

1.2.4 m-Factors-at-a-Time (mFAT) Designs

In contrast with methods of conditioning on the main effects, the m-factors-at-a-time design was presented by Larsen (2023) as a method of balancing the variance and bias intrinsic to an OFAT experiment without the cost of running a TvC analysis:

Defintion: A m-factors-at-a-time (mFAT) design is any factorial design that replicates at least one treatment where m factors are set to $j_l = 1$ and all other replicated treatments have m or fewer factors set to $j_l = 1$. That is, every treatment combination replicated under an mFAT design satisfies $\sum_{l=1}^{k} j_l \leq m$ with at least one treatment combination where $\sum_{l=1}^{k} j_l = m$.

The OFAT design is a special case of the mFAT in which $m = 1$. An mFAT design can replicate up to $p = \sum_{l=0}^{m} \binom{k}{l}$ treatment combinations but it can be less than that. Let $\mathbb{T}_m = (j_1 j_2 \cdots j_k)_{i=1}^{p}$ denote this largest collection of treatment combinations, so $|\mathbb{T}| = p$. Table 1.1 lists various mFAT designs and the treatments examined in each for $k = 3$.

Table 1.1: Examples of mFAT designs. Rows 1 - 4 present examples of the treatment combinations under mFAT designs considered plausible in this paper. Combinations of treatments like the ones in row 5 are not considered.

OFAT	(000)	(100)	(010)	(001)				
2FAT	(000)	(100)	(010)	(001)	(110)	(101)	(011)	
2FAT	(000)	(100)	(010)	(001)	(110)			
3FAT	(000)	(100)	(010)	(001)	(110)	(101)	(011)	(111)
3FAT	(000)	(100)	(111)					

In this paper we only consider an mFAT design, d, that replicate treatments in $\mathbb{T}_m$. The cell means model under such a design is

$$y = X_d \mu_{\mathbb{T}_m} + \epsilon \tag{1.16}$$

where $X_d \in N \times p$ and $\mu_{\mathbb{T}_m} \in \mathbb{R}^{p \times 1}$ are submatrices of the corresponding items in (1.8) following $\mathbb{T}_m$. For example, if $k = 3$ then there is a potential $2^3 = 8$ possible treatment combinations. If we are only interested in measuring main effects, we would use an OFAT design so $m = 1$ which replicates $p = 4$ treatments: 000, 100, 010 and 001. This results in X_d being reduced to 4 columns and the vector $\mu_{\mathbb{T}_m}$ only lists the means for those four treatments.

The $\mathbb{T}_m$ subscript is henceforth dropped from μ to simplify notation allowing the p BP factorial effects of interest to be denoted by $\theta = H\mu$, where we have also dropped the BP notation for simplicity. For a given design d, the variance of the corresponding effect estimates is $Var_d(\hat{\theta}) = H^T R_d^{-1} H$, where R_d is the diagonal matrix of replications.

1.2.5 Optimal mFAT designs for BP

Typically, determining a design, d, would require identifying both the treatment combinations to replicate, and how often they should be replicated. However, we simplify the problem by requiring that every treatment in $\mathbb{T}_m$ is replicated at least once to ensure all BP m-factor interactions and lower-order interactions are estimated unbiasedly. Having conditioned on the support for the designs, the next step is to calculate the optimal replicates per treatment in the support. To start, we consider the optimal replicates to be those that simultaneously minimize the variances of the baseline effects of interest.

Simultaneously minimizing the variance of the BP effects of interest is likely impossible, especially since the minimum possible variances are different across the BP effects. An alternative is to use an optimal design criterion that pushes all the effects' equally towards their minimum possible variance. Standard approaches to this problem form the basis of traditional optimal design theory and are detailed in Pukelsheim (2006). Optimality criteria based on eigenvalues are popular methods from this theory and part of a broader class of weighted optimality criteria (Stallings and Morgan 2015) that target weighted effect variances for any estimable function $h^T\mu$,

$$Var_{dW}(h^T\mu) = (h^T W^{-1} h)^{-1} Var_d(h^T\mu), \tag{1.17}$$

where W is any $p \times p$, symmetric, positive-definite weight matrix. The weights are assumed to be defined by the practitioner and are not design-dependent. Typically, each estimable function is assigned a weight w_i such that $D_W = diag(w_1, ..., w_p)$ and

$$W = (I - Q) + H^T D_W H, \tag{1.18}$$

where Q is a projection matrix onto the row space of H. Weighted optimality criteria are analogous to their non-weighted counterparts, defined as $\Phi(C_{dW})$, where $C_{dW} = W^{-\frac{1}{2}} C_d W^{-\frac{1}{2}}$ is the weighted information matrix. When W is defined as in equation (1.18), the average of the inverse eigenvalues of C_{dW} is proportional to the average variance of the p estimable functions weighted by D_W. Stallings and Morgan (2015) refer to this quantity as the weighted-A criteria, A_w, which is used to determine optimal mFAT designs in the heuristic presented by Larsen (2023).

Baseline effects under an mFAT design correspond to estimable contrasts of the form $h_j^T \mu$. It follows that for a given design d, the j-th baseline effect has a variance proportional to $\sum_{i=1}^{p} \frac{h_{ij}^2}{r_{di}}$, where $r_{di} > 0$ denotes the number of replications of treatment i. Stallings and Morgan (2015) established the following for approximate CRD designs.

Theorem 1. *Let $w_i^* = \Sigma_{j=1}^{p} w_i h_{ij}^2$. Then the approximate design d^* with replications*

$$r_{d^*i} = N \frac{\sqrt{w_i^*}}{\Sigma \sqrt{w_i^*}}$$

minimizes the average weighted variance of the p baseline effects of interest and is therefore A_w optimal.

Stallings and Morgan (2015) also show that, for a given A_w optimal design of N runs, any A_w optimal design consisting $N^* > N$ runs can be found with the following recursive relation.

Theorem 2. *Given an N-run, A_w optimal design with replication numbers $r_1^*, ...r_p^*$, the*

$(N+1)$*-design produced by adding one run to any of the* p *treatments satisfying*

$$max \frac{w_i^*}{r_{i,N}^*(1+r_{i,N}^*)}$$

with the other $p-1$ *treatments unchanged, is* A_w*−optimal for* $N+1$ *runs.*

It follows that to create an optimal A_w mFAT design with N runs, one can begin with the trivial A_w optimal design with $r_{d*i}=1$ for all treatments and recursively apply Theorem 2. Table 1.2 illustrates this process for a 2FAT design estimating all main effects and two two-factor interactions where $k=3$ and $D_w=I$. This example demonstrates how the A_w optimal designs lend themselves to baseline parametrization as compared to balanced designs for orthogonal effects. These designs replicate treatments whose means are required to estimate the most baseline effects, prioritizing the baseline treatment and then the 1FAT treatments, leaving fewer runs for treatments whose means are used to calculate fewer effects.

Table 1.2: A_w-optimal 2FAT designs, $k=3$

	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
000	1	2	2	3	3	3	4	4	4	4	4	4	5	5	6	6	6	6
100	1	1	2	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4
010	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	4	4	4
001	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	4
110	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	3	3
101	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2

The composition of D_w is customized for each experiment to place the desired emphasis in the A_w optimal design on shrinking the variance of a given effect estimate toward its minimum. This section assumes equal importance of effect estimates when minimizing their variances. When orthogonal effects are equally weighted, the optimal choice for D_w is I, with all $w_i = 1$.

However, baseline effects do not meet this criteria as the interaction effects, lower order effects and the control all have different minimum variances. In this situation, the A_w criterion tends to emphasize the effects with the largest minimum variances (Allen-Moyer and Stallrich 2022). To encourage this further, Allen-Moyer and Stallrich (2022) suggest choosing values for D_w that are inversely proportional to the minimum variances.

Continuing the example in Table 1.2, this means defining $D_w = diag((N-3)^{-1}, 4(N-2)^{-1}, 4(N-2)^{-1}, 16N^{-1})$. For a given mFAT effect for a fixed N, its minimum possible variance is,

$$\mathrm{Var}_*(\hat{\theta}_{j_1 j_2 \cdots j_k}) = \frac{2^{2m}}{N-(p-2^m)}, \tag{1.19}$$

which can be used to define $D_w = diag(\frac{1}{var_*(\theta_{j_1 j_2 \cdots j_k})})$, Table 1.3 applies the standardized weighting to the previous example.

Table 1.3: Standardized A_w-optimal 2FAT designs, $k = 3$

	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
000	1	2	3	3	3	3	4	5	5	6	6	6	7	7	7	8	8	8
100	1	1	1	2	2	2	2	2	3	3	3	3	3	3	3	3	4	4
010	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	3	4
001	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	3
110	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
101	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2

The table demonstrates that optimal standardized A_w designs prioritize replicating the treatments that appear most frequently in low order effects, or effects corresponding to a lower minimum variance, more than the non-standardized designs. While this weighting comes at the cost of replicating treatments that appear most commonly in higher order effects less frequently, the impact of this while examining higher-order interactions is not substantially noticeable. In contrast, the standardized designs push the variances of the lower order effects considerably closer to their minimums.

The optimal mFAT designs address many of the systemic issues which are not adequately addressed in current baseline parametrization research. Namely, by building from both the OFAT model and baseline parametrization, it provides a way to achieve unbiased estimates of all effects of interest without requiring either at least one replication of all 2^k settings, or higher-order effects to be negligible. At the same time, it works to limit the variance of estimated effects.

This design can also be deployed in a sequential framework, which can be used to mimic the real-time nature of many OCEs, by assuming experimental units arrive sequentially in blocks of size $c \geq 1$. Each time step would cause the baseline to be updated to the best performing variant once that has been reliably determined, which maximizes reward while still prioritizing baseline effect analysis. This allows for the sequential randomization of experimental units such that: (1) unbiased, minimal variance estimates of the baseline effects are ensured, provided no additional assumptions are made regarding effect structure, and (2) the best performing treatment combination is safely identified. “Safe" is defined as minimizing the replication of poorly performing treatments, if they are replicated at all. One aspect of safe exploration is careful definition of $\mathbb{T}_m$. The other aspect is the use of sequential, adaptive strategies to determine the selection of subsequent sets of treatments. Larsen (2023) proposes a heuristic providing these features which uses the properties of the sequential frameworks and optimal mFAT designs under baseline parametrization to flip the paradigm of recent research on how best to approach minimizing bias and variance in baseline parametrization estimates. While the rest of this paper focuses only on the OFAT special case, we present the mFAT designs as a possible extension of this work.

1.2.6 Sequentially Weighted mFAT

The SW mFAT method is a heuristic proposed by Larsen (2023) that addresses the challenges of subset learning and optimization stated in the beginning of this paper. It deploys the mFAT design through a sequential framework such that each stage of the process can reliably identify the best performing variant while still prioritizing baseline effect analysis. Upon completion of each stage, the heuristic calculates whether the baseline should be updated and thus which treatments to discard. If so, the optimal design for the next stage is calculated based upon the new support and all information collected thus far.

Within each stage, experimental units arrive sequentially in fixed block sizes of c, which sum to a total of n units. The value of n is calculated based on power curves determined by the experimenter's selection of the pairwise threshold for significance, α, and power, β. These curves are generated using a t-distribution with non-centrality parameter $\frac{|\delta|}{SD(\hat{\theta}_{main})}$ and degrees of freedom set to the number of desired contrasts. Here $SD(\hat{\theta}_{main})$ refers to the standard deviation of a main effect under baseline parametrization assuming a non-sequential experiment with n runs where treatments were replicated under the optimal replicates found in Theorem 1.

The first block is allocated based on an unweighted standardized A-criterion model. The subsequent blocks use the cumulative p-values of the estimated effects to weight the allocation of experimental units. This is implemented by multiplying the weights in D_w (inverse minimum effect variances) by the current p-values of the estimated effects,

$$PD_w = diag(\frac{p_{j_1,j_2,\ldots,j_k}}{\text{Var}^*(\hat{\theta}_{j_1,j_2,\ldots,j_k})}) \quad (1.20)$$

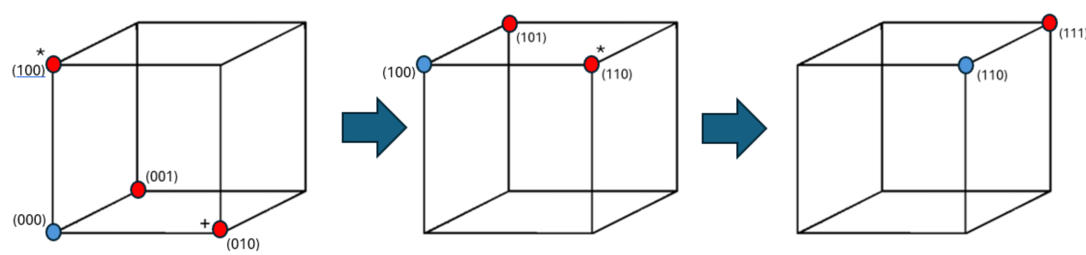
where P is a $p \times p$ matrix with the p-values on the diagonal and zeros elsewhere. By using the p-values in the weighting matrix, the model is oriented to spend more time exploring treatments involved in effects that are currently statistically insignificant.

After each block, effect estimates are updated and the previous design is augmented with the new units. A new optimal design d_b is found according to the theorems in Section 1.2.5 with a weight matrix $P_{b-1}D_w$, where P_{b-1} is the diagonal matrix of p-values from the previous block. The new experimental units are randomized according to d_b and a new P_b is computed.

The stage can end in a variety of ways. If all elements of the diagonal of P_b are less than some α, meaning that all the effects are statistically significant, the experiment can be ended immediately, prior to the allocation of all N. Barring this, the fully powered stage will run to completion. Once sufficient data has been collected, the experimenter may choose to (1) end the experiment and declare the best performing treatment as the optimum or, if there are no significant improvements found, declare the baseline as the selected treatment, (2) update the baseline to the observed best performing treatment and conduct an additional SW mFAT stage or (3) calculate extrapolating estimates of the unobserved treatment means and update the baseline to the predicted best treatment.

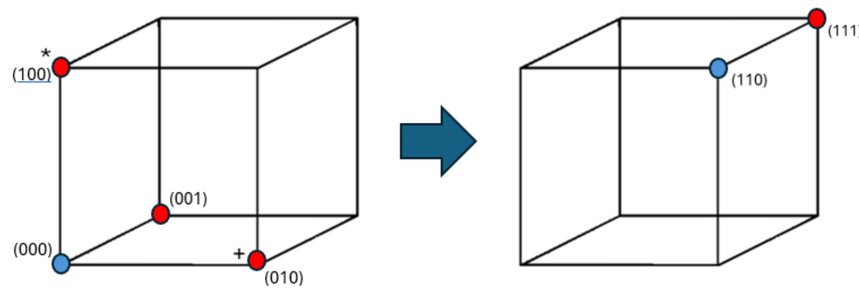
The first option of continuing the experiment by updating the baseline to the observed best performing treatment and conducting an additional SW mFAT experiment works much like an OFAT design. After the first stage is completed, another stage is conducted, this time considering all treatments defined with respect to the new baseline. The process is repeated until all relevant factor combinations have been exhausted. See Figure 1.4 for a graphical explanation of this option. The strengths of this approach lie in the fact that there is no guesswork regarding the interaction effects, as it takes the safest route by always choosing an observed treatment. However, the sample size requirements increase proportional to the number of treatments replicated and there is a risk in replicating a larger number of treatments.

Figure 1.4: SW mFAT updating to the observed best performing treatment. “$*$" indicates the best performing treatment in that stage and "+" indicates a treatment which improved on the baseline.



The second option for continuing is via extrapolation. This method also requires running another stage on an updated baseline. However, the new baseline is selected as the combination of all significantly positive treatments. For example, if the first stage found that $\hat{\mu}_{000} = 3$, $\hat{\mu}_{100} = 3.2$, $\hat{\mu}_{010} = 2.9$ and $\hat{\mu}_{001} = 3.1$ then the new baseline would be 101. See Figure 1.5 for a graphical explanation of this option. This method requires a smaller sample size than the first approach, however it also depends on assumptions that the effects of interaction terms are greater than or equal to 0. If those assumptions are met, this options accomplishes the same goals as the previous option with a smaller sample size.

Figure 1.5: SW mFAT updating to an extrapolated baseline. “$*$" indicates the best performing treatment in that stage and "+" indicates a treatment which improved on the baseline.



SW mFAT successfully provides a way to balance both subset learning and optimization. By flipping the common paradigm for methods that identify optimal treatment combinations, it minimizes replication of significant treatment combinations to focus exploration on the rest of the treatment set. In this way, it provides more opportunity to discover information regarding treatments with potentially very small but present effect sizes. However, it still maintains enough replication of these significant treatments so as to correctly identify the optimal treatment in each stage. Further, by addressing the problem in stages, this method allows for very safe exploration of the treatment set.

1.3 Pitfalls of Sequentially Weighted mFAT

While SW mFAT took many steps forward in terms of not just safely identifying optimal treatment combinations in circumstances unique to OCE, it is not robust to all possible scenarios an experimenter may be in. We discuss these pitfalls below.

1.3.1 Type I Error and Block Size

The first drawback to the SW mFAT method lies in the definition of the decision-making rule to stop the stage before using all n. Early termination of the stage is contingent on every treatment combination returning a significant effect size, meaning the easier it is for an effect size to be found significant, the easier it is to end the experiment early. SW mFAT uses a pairwise type I error rate to make this determination. This results in a perpetually compounding error rate as the number of treatments, k, increases. Since the strength of this method lies in managing circumstances with higher k, this increasing error rate is particularly problematic. As k increases, stages will be more likely to erroneously end early. This can result in both providing incorrect information for assigning the best baseline for future stages, as the significance of the effect size measured would have been incorrectly calculated and a loss of information on all treatments that would have been collected had the stage been run its entirety.

Another pitfall in the method's design decisions is the determining of block sizes. While the simulation studies provided in Larsen (2023) run with a block size of 25, there is no guidance on how to determine the bounds of an appropriate block size for a given experiment. The correct calculation of block size is key to the method's success. Smaller blocks can compound the over-replication of the control. This is exemplified in Table 1.3 which demonstrates the optimal allocation for each block size. When the block is size 6 for a $k = 5$ experiment, the only possible allocation is for each treatment to receive one experimental unit. At block size 7, the method will allocate the additional unit to the control as the impact of minimum variance adjustments on the allocation is substantially higher than that the weight of the p-values. As such, the control ends up with double the experimental units as compared to each other treatment in each block. This compounds with each block to create an over-replication of the control. The impacts of this over-replication are discussed in depth later in this section. Larger blocks may not leave enough time for the method to learn, thus potentially failing to identify treatment combinations with significant but small effect sizes. The use of blocks was motivated by Haizler and Steinberg (2021) and is clearly a reasonable choice in the environment in which we are testing via the extension of the optimal A_w design theorems in Section 1.2.5 (Larsen 2023). However, without clarity on how to determine an ideal block size the method leaves ample opportunity for poor decisions.

1.3.2 Over-replication of the Control

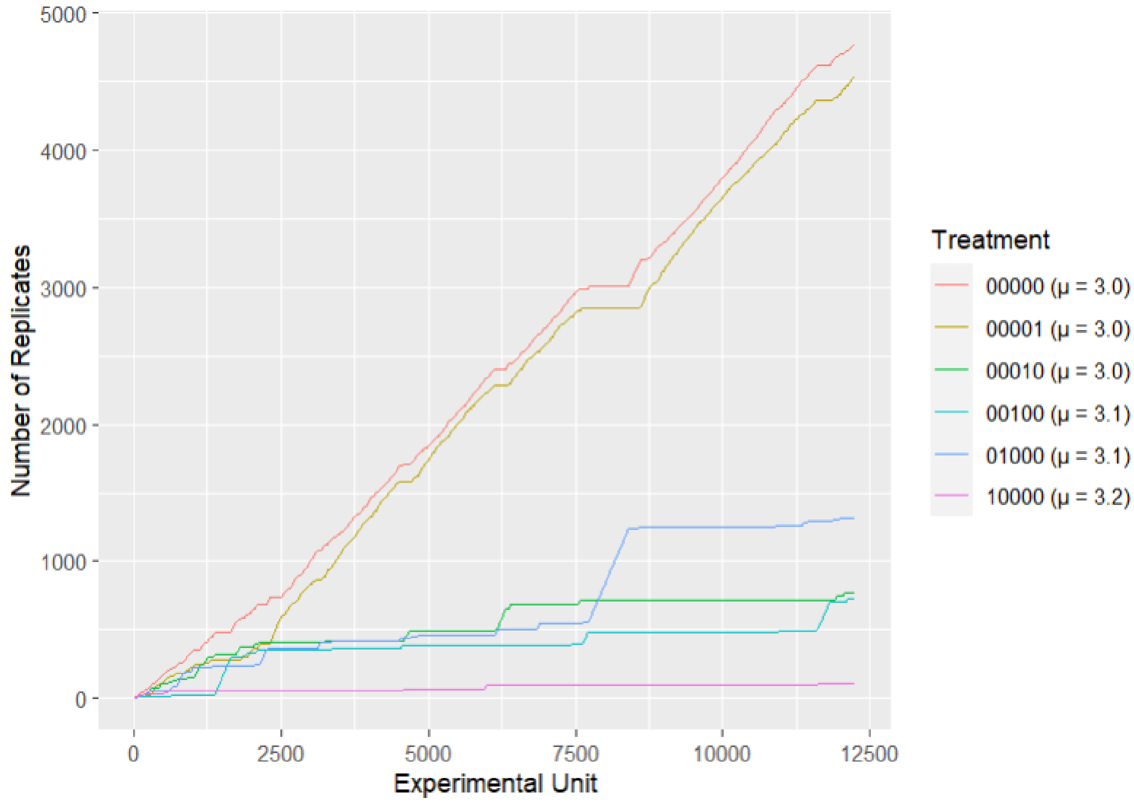
The second drawback is the over-replication of the control. The purpose of this method is to identify the treatment combination, if there is one, which results in the largest effect size. This means that resources do not need to be allocated toward estimating the control treatment mean outside of providing enough information to calculate said effect sizes. SW mFAT prioritizes lowering the variance of the estimate of the control treatment mean in addition to this, which produces an over-allocation of resources to investigating the control. The over-weighting is compounded by the method of experimental unit allocation. This occurs as a result of the minimization of the variances examined in section 3.1 in the creation of the weight matrix. Table 1.4 lists the minimum variances for various treatment combinations for $k = 2$ under baseline parametrization.

Table 1.4: Variance calculations for baseline effects for $k = 2$. Regardless of the number of treatments examined, the control will always have the lowest minimum variance.

	Variance	Min. Variance
$\hat{\theta}_{00}^{BP}$	r_{00}^{-1}	$\frac{1}{N-3}$
$\hat{\theta}_{10}^{BP}$	$r_{00}^{-1} + r_{10}^{-1}$	$\frac{4}{N-2}$
$\hat{\theta}_{01}^{BP}$	$r_{00}^{-1} + r_{01}^{-1}$	$\frac{4}{N-2}$
$\hat{\theta}_{11}^{BP}$	$r_{00}^{-1} + r_{10}^{-1} + r_{01}^{-1} + r_{11}^{-1}$	$\frac{16}{N}$

By definition, the control will have a lower minimum variance than any of the contrasts being examined due to having fewer terms. This results in the allocation method favoring the control in an effort to lower the overall variance most effectively. The focus placed on the control can be seen in Figure 1.6.

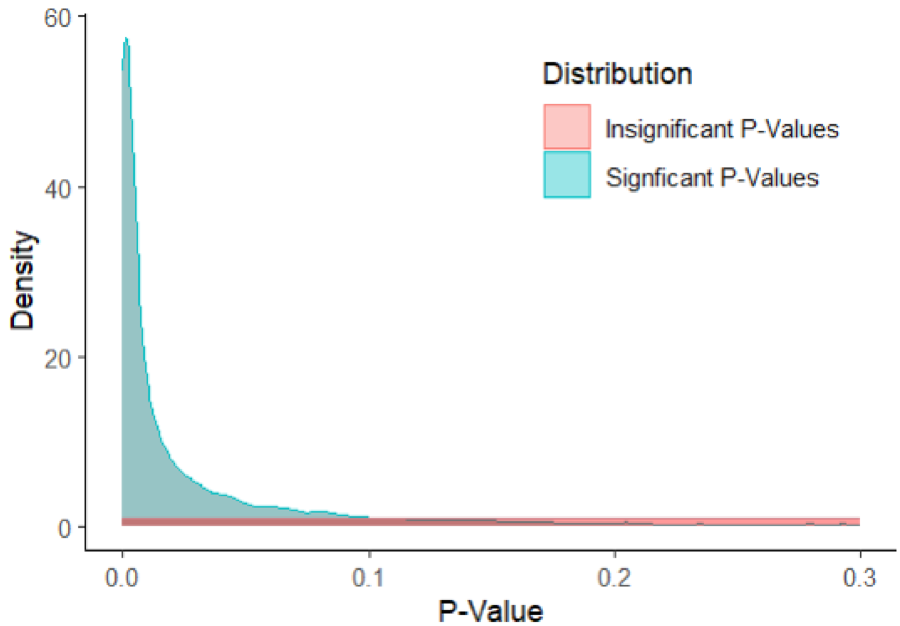
Figure 1.6: Visualization of SW mFAT design. The control is replicated unnecessarily when the experimental units could be used to better explore other treatments more equally.



1.3.3 Weighting with p-values

The third area to be addressed is the method's reliance on the fallacy that p-values are a measure of the strength of significance and insignificance. SW mFAT uses p-values as proportional weights for allocation instead of simply as indicators of significance. The root of this error lies in the duality of the distribution of p-values. When a treatment has no significant difference from the control the corresponding p-values will be distributed $U(0,1)$. However, when a treatment does have a significant difference from the control the corresponding p-values will be heavily weighted towards 0. Figure 1.7 displays examples of both distributions.

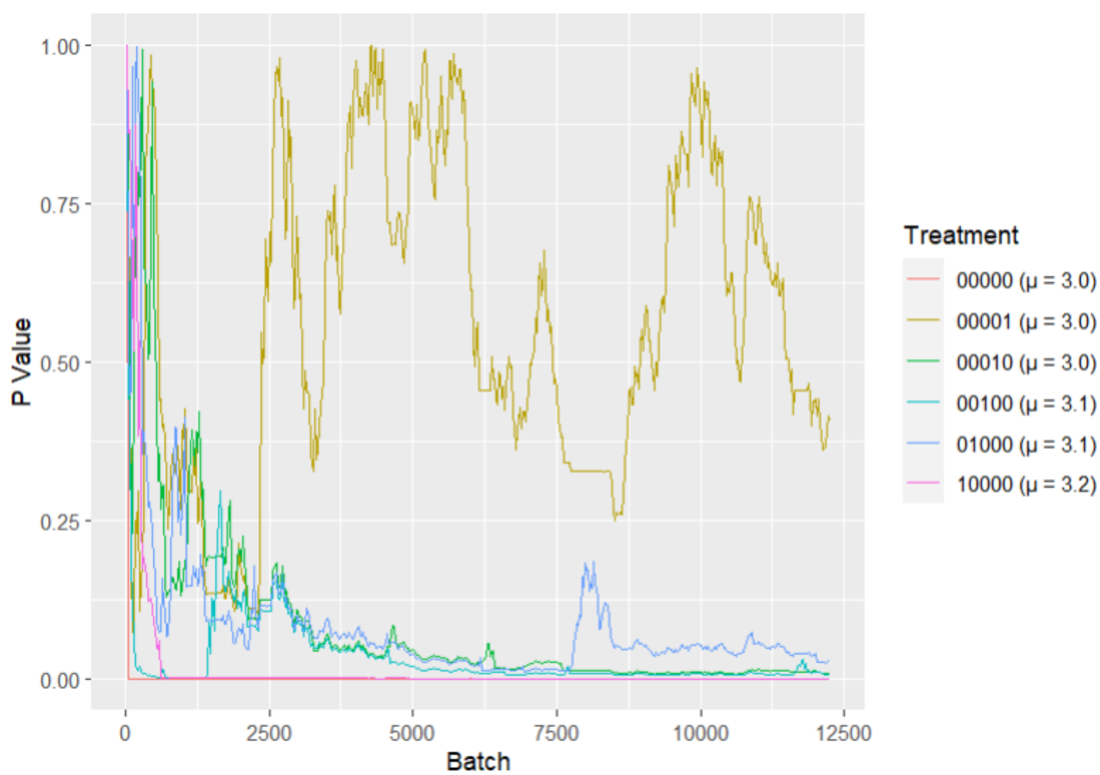
Figure 1.7: Comparison of the distributions of p-values. The PDF for $U(0,1)$ was plotted as the insignificant p-value distribution against significant p-value data generated via simulations. The plot was truncated at p-value = 0.3 for readability.



Thus, while it is tempting to simply state that the lower a p-value is, the more significant it is, this interpretation is not uniformly true. It is incorrect to imply that a p-value of 1 is particularly insignificant as compared to a p-value of 0.2, as they are both simply not statistically significant. The SW mFAT, by using the exact p-values as weights for allocation, indirectly relies on this faulty assumption.

The impact of this can be seen in two phenomena. The first occurs when an observed p-value for a baseline main effect is close to 0. In the case of an OFAT design targeting baseline main effects, this results in the method rarely replicating the corresponding treatment. This is intentional in the case of a significant treatment effect size. However, since this can happen randomly for treatment combinations with insignificant effect sizes, it is possible for one of these treatments to be falsely identified as significant. Some steps were taken to minimize the impact of random low p-values early in the stage, such as keeping all treatments in the model even after they are found significant. However the weighting for those treatments becomes so low the method effectively never allocates to those treatments again, which can lead to type I errors. Figure 1.6 demonstrates an example of these circumstances. The corresponding p-values used for weighting for each block can be seen in Figure 1.8. Treatment 00010 should have been equally as explored as 00001 seeing as they both have an effect size of 0, however due to a p-value returning low early in the stage, treatment 00010 was not explored further since it was being weighted as though it had a significant effect.

Figure 1.8: Visualization of SW mFAT Weighting p-Values. The treatment 00010 randomly returned a low p-value which never recovered as the treatment was not replicated.



The second potential issue also relates to non-ideal replication patterns, though the consequences are not as extreme as the first. Continuing with the same stage described previously, consider the situation in which treatment 00001 happened to return a p-value close to 1 while 00010 returned one closer to 0, though evidently not significant. While treatment 00010 still receives experimental units allocated to it, an improvement over the previous example, the allocation would still favor 00001, thus leading to arbitrary unequal exploration. Furthermore, because the method uses cumulative knowledge to create the p-values for each block, the discrepancy is perpetuated unless corrected by another random chance p-value. Figures 1.9 and 1.10 demonstrate an example of the replication and corresponding p-value weights for this situation.

Figure 1.9: Visualization of SW mFAT design. Treatment 00001 is explored more than 00010 due to the random pulls of p-values for each treatment.

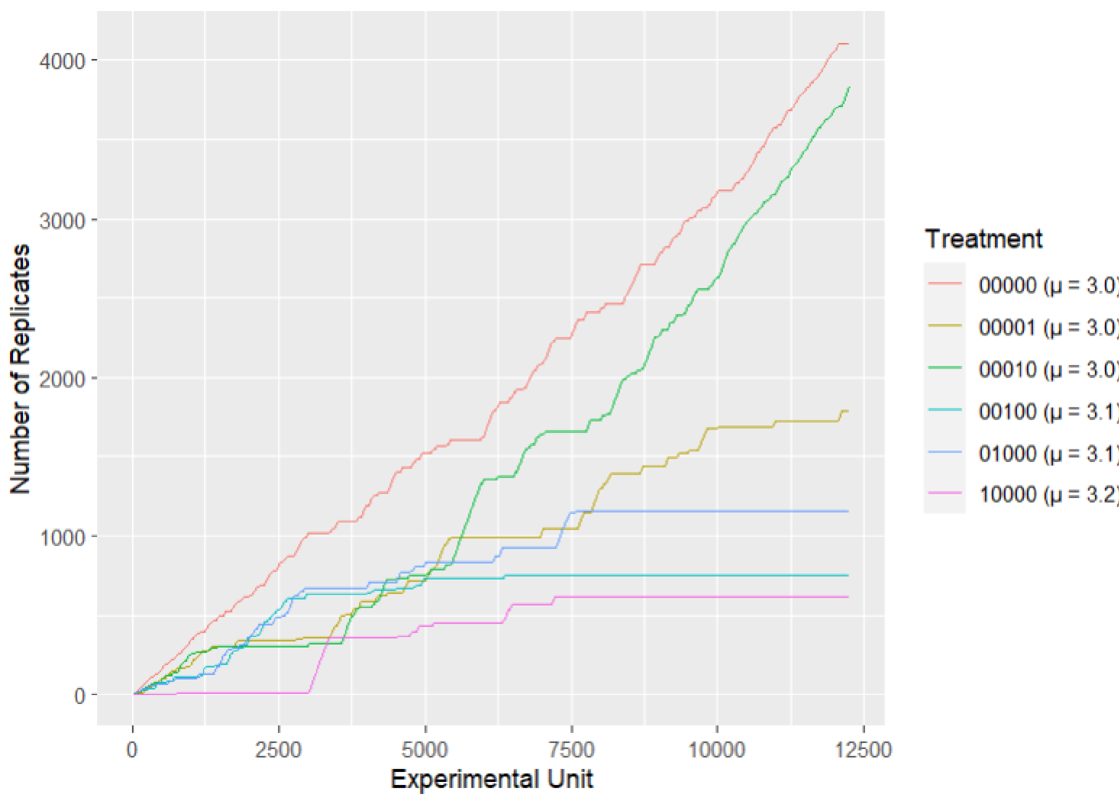
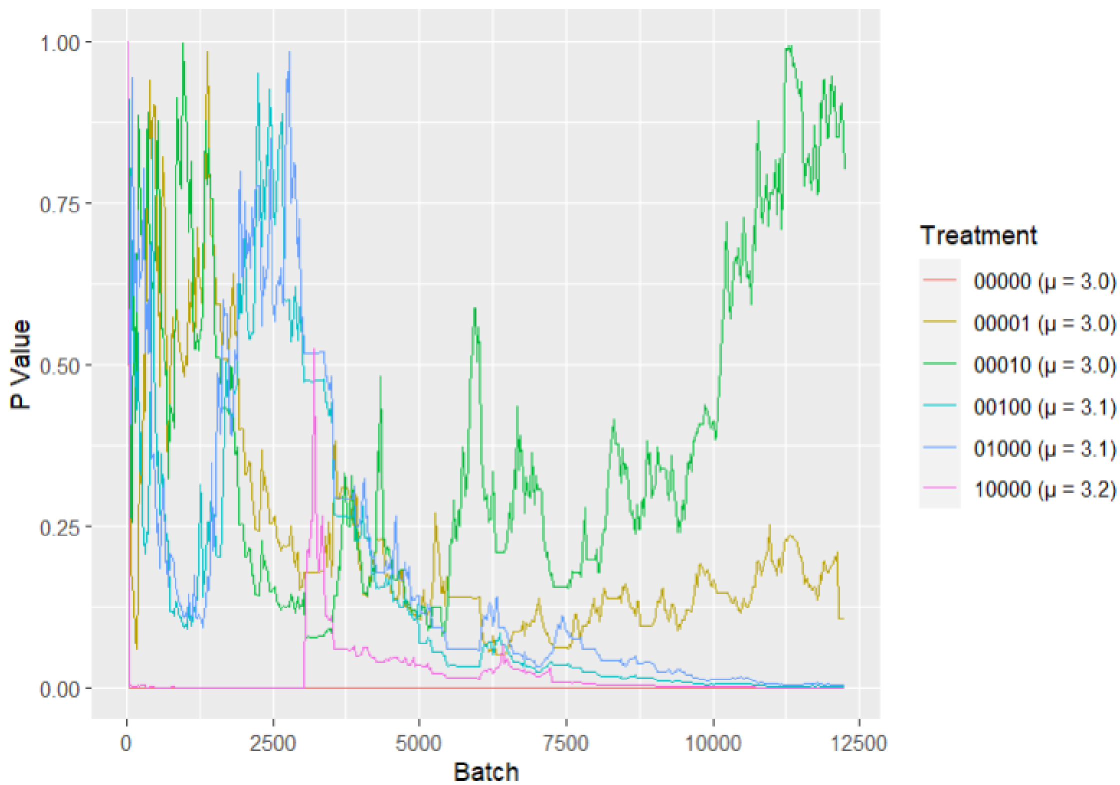


Figure 1.10: Visualization of SW mFAT Weighting p-Values. The treatment 00010 randomly returned a low p-value as compared to treatment 00001, resulting in unequal exploration. Recovery did not occur as it would rely on another random and equally extreme p-value being added to the cumulative weight.



This inequality in exploration actively undercuts the method's second goal of identifying further information for the practitioner about the treatments being explored.

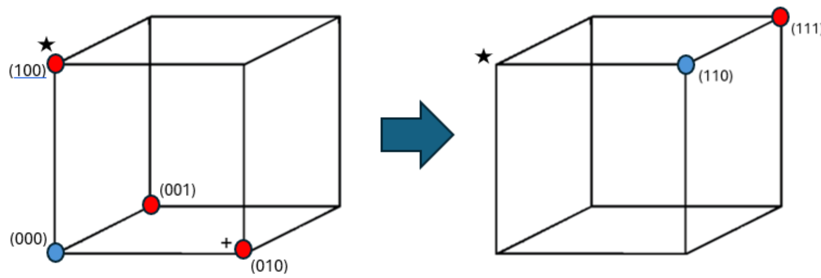
1.3.4 Extrapolation Assumptions

As discussed earlier in this section, one of the possible continuations after the first round of SW mFAT is to combine all treatments from the round with significant positive effect sizes to determine a new baseline. Larsen (2023) notes that the extrapolation process requires the assumption that the effect sizes are purely additive, or in other words that the interactions between all the chosen treatment combinations are 0. In order to further develop this proposal, we require additional examination of how the new extrapolated baseline is determined.

The extrapolated baseline is defined as the combination of all treatments with significantly positive effect sizes. In this way, the experimenter can skip the step by step addition of new treatments to the control and move directly to a combination that is strictly superior to the control and all stage one treatments. For example, consider a stage 1 run for $k = 3$ treatments that resulted in $\hat{\mu}_{000} = 3.0$, $\hat{\mu}_{100} = 3.2$, $\hat{\mu}_{010} = 3.1$, and $\hat{\mu}_{001} = 3.0$. The extrapolated baseline would be 110 as seen in Figure 1.5. When working under the assumptions proposed by Larsen (2023) the interaction between the two treatments is 0, meaning $\theta_{110} = 0$. As such, $\hat{\mu}_{110} = \hat{\theta}_{100} + \hat{\theta}_{010} + \hat{\theta}_{000} = \hat{\mu}_{100} + \hat{\mu}_{010} - \hat{\mu}_{000} + 0 = 3.3$. So long as the assumption that there is no interaction between the initial treatments holds, extrapolation in this manner will always be safe for the practitioner.

However, in practice, the practitioner will not know if these assumptions are met. As such, there is no clear guidance on how to go about using this option and the experimenter must choose based on what, if any, background information they may have. This can lead to either an unnecessary round of SW mFAT if the experimenter incorrectly asserts that there are no interactions. Alternatively, falsely claiming the assumptions to be true can lead to never identifying the best possible treatment combination of those being examined, as demonstrated in Figure 1.11. The consequences of this are further explored in the simulation study examined in Section 1.5.3.

Figure 1.11: Visualization of SW mFAT when extrapolation assumptions are not met. "⋆" indicates the best performing treatment in the experiment. Due to the assumption that interaction between 100 and 010 is at least 0, the heuristic bypasses the optimal treatment combination.



1.4 Improved SW mFAT

This section proposes an algorithm that addresses the critiques of the SW mFAT method discussed in the previous section. The algorithm aims to maintain the lower regret as compared to MAB models and the ability to safely explore the factorial space which the SW mFAT demonstrates while concurrently minimizing error and arbitrary decisioning.

1.4.1 Bonferonni Correction

In response to the first drawback discussed in Section 3.1 concerning the control of type I error regardless of k, we propose implementing a Bonferroni correction and letting the family-wise type I error remain α, and each pairwise value be α/k. This will prevent the increase in type I errors as k grows. Table 1.5 demonstrates the difference in error rates between the two methods.

Table 1.5: Expected Type I error rates for each value of k for SW mFAT and Improved SW mFAT

Type I Error Rate		
		Improved
k	SW $mFAT$	SW mFAT
3	14.3%	5.0%
5	22.6%	5.0%
7	30.2%	5.0%

An additional guard on type I error is discussed in Section 4.3 as it is intrinsic to the allocation method proposed. This change in the value of α also impacted sample size calculations. In order to maintain the same power level with a more restricted type I error, the sample size needed to increase. The sample size for the first stage of experiments with $k = 3, 5,$ and 7 are shown in Table 1.6 for both methods.

Table 1.6: Sample sizes for each value of k for Stage 1 of SW mFAT and Improved SW mFAT methods

Stage 1 Sample Size		
		Improved
k	SW $mFAT$	SW mFAT
3	6,034	7,855
5	8,270	12,260
7	10,408	16,598

1.4.2 Adjusting out the Control in A_W-criterion

When examining the drawback of over-replication of the control, described in Section 3.2, it becomes clear that while a compound issue in SW mFAT, the solution is quite elegant. By recognizing that the goal of this method, and in fact any OCE, is not to correctly predict the value of the control, but to identify which treatment(s), if any, show a significant and positive effect size, we can simply remove the control from the list of treatments being estimated.

As such, this heuristic removes the row concerning only the control from the H matrix and corresponding list of weights, ensuring that each allocation to the control is only to improve the estimates of the effect sizes. In the case of an OFAT design targeting the k baseline main effects, this causes the the W matrix from Section 2.6 to be calculated as:

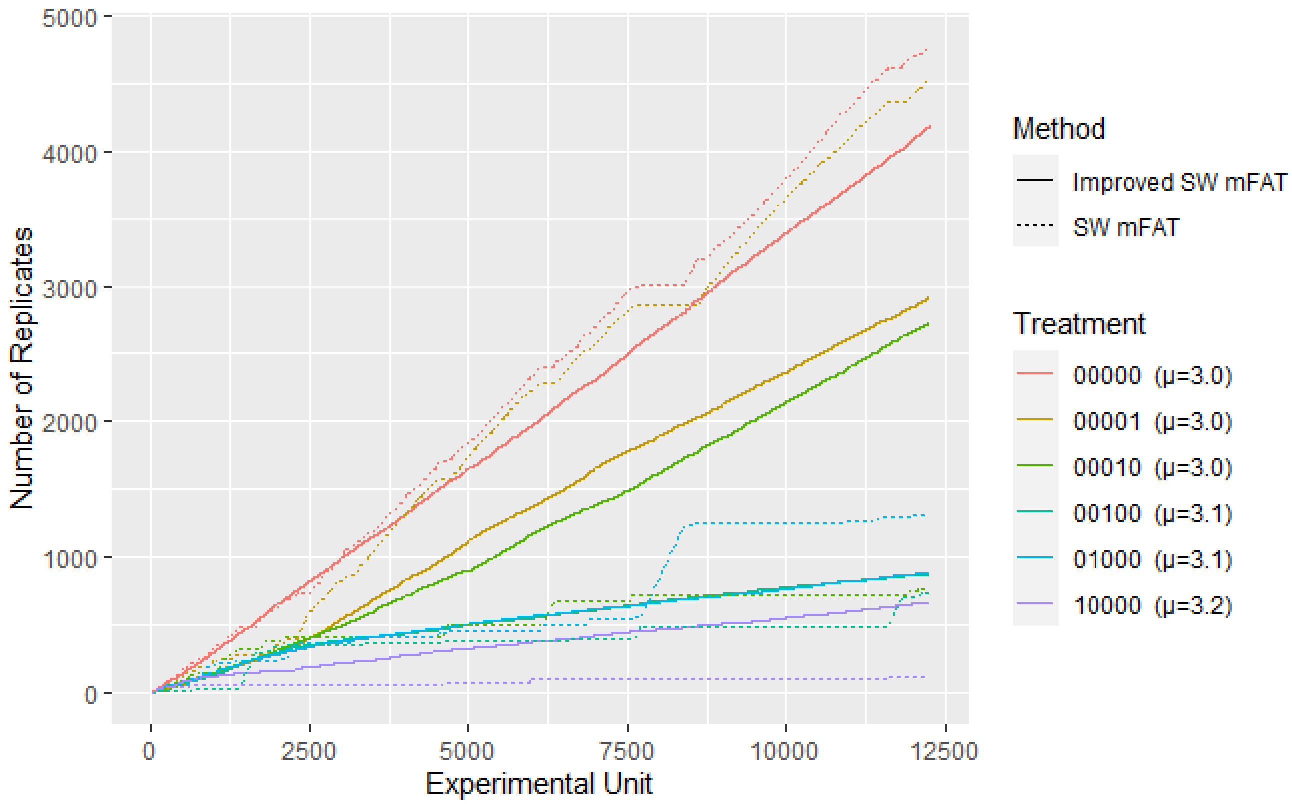
$$W = H^T_{k+1\times k} D_{k\times k} H_{k\times k+1} + \frac{1}{k+1} 1_{k+1} 1^T_{k+1} \tag{1.21}$$

Due to all baseline main effects having the same minimum possible variance, we do not require the inverse minimum variance adjustment and set $D^*_w = I$. As such

$$PD^*_w = diag(p_{j_2,j_3,\ldots,j_k}) * I \tag{1.22}$$

which results in the weight matrix simply being $diag(p_{j_2,j_3,\ldots,j_k})$. As a result, the method maintains its focus on the priority, which is identifying the treatment combination which results in the largest effect size. The control remains well estimated, as the allocation method still favors it due to the increase in information it provides to all effect estimates. Figure 1.12 exemplifies the decrease in control replication and more equal replication of treatments with insignificant effect sizes.

Figure 1.12: Visualization of Improved SW mFAT design as compared to SW mFAT. The control is now replicated fewer times allowing for better and more equal of exploration other treatments.



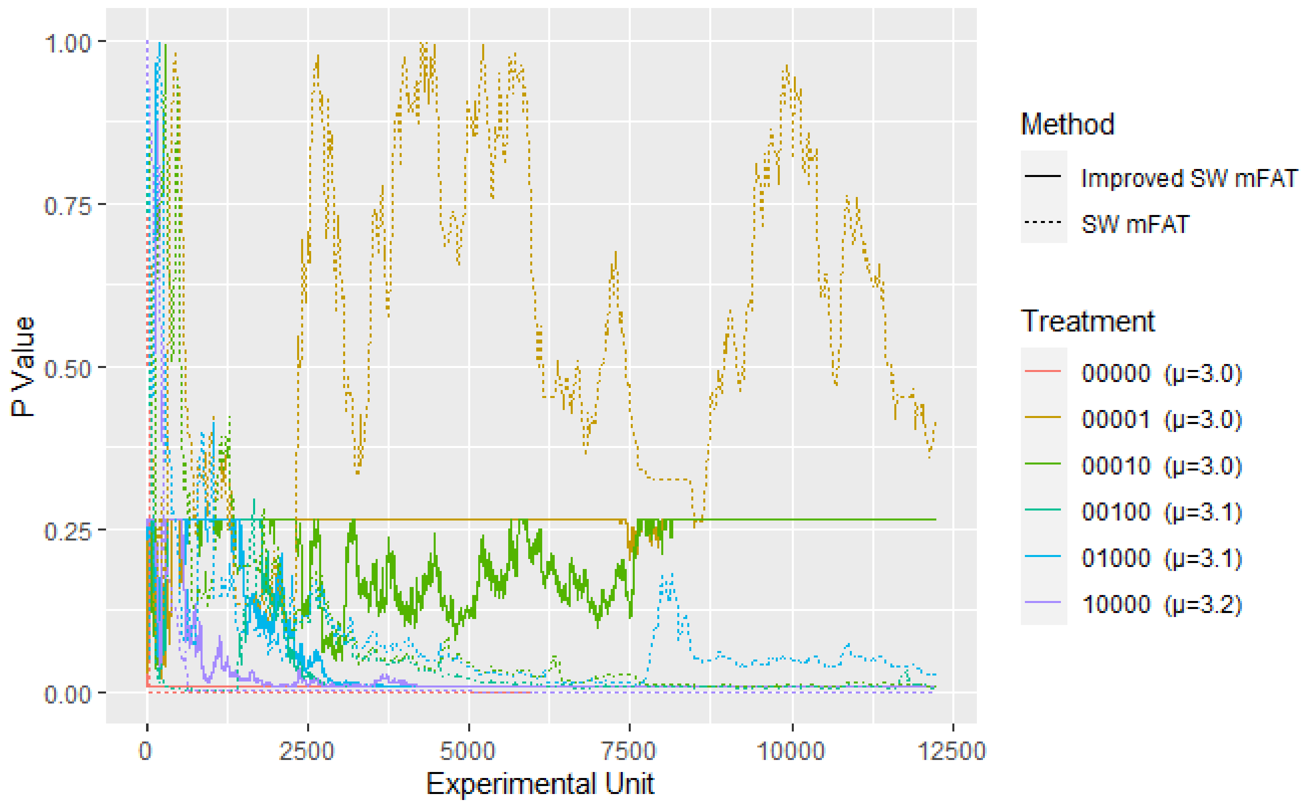
1.4.3 Improved Stage Allocation

The next improvement to SW mFAT we propose is the implementation of a new allocation algorithm. This addresses aspects of both the drawbacks in experimental design decisions (Section 3.1) and allocation within stages (Section 3.3). The basis for this proposal is the missing guidelines for the calculation of appropriate block sizes. While Larsen (2023) makes a compelling argument as to the the sequential nature of the process lending itself well to using blocks, it also allows us to simplify the process to allocate each experimental unit as it comes in and thus eliminate the need to use blocks at all. However, this effectively makes each block size equal to 1 and as discussed in section 3.1, SW mFAT does not perform well with very small block sizes. As such, we propose a new allocation algorithm.

For a OFAT design, the first $k+2$ experimental units are randomly assigned to so one goes to the test treatment combinations and two go to the control, so that each has a starting data point from which to calculate p-values. Then upon each subsequent experimental unit's arrival, we calculate the most recent p-values for each treatment and the control before capping them using a minimum and maximum p-value. The minimum p-value is set to α, after the Bonferroni adjustment. To find the maximum p-value, a Monte Carlo simulation is run to find the distribution of p-values for a treatment with the minimum effect size being measured, δ. We set the upper bound to be the 0.99 quantile of this distribution. In this way, the vast majority of p-values returned by treatments with significant effect sizes are left as is, allowing the algorithm to weight them appropriately lower than non-significant treatments which we wish to explore more. Conversely, the set of treatments with truly insignificant effects sizes which would frequently return p-values greater than this quantile, will be assigned the same weight (i.e. the quantile value) and so will be equally explored.

Figure 1.13 demonstrates the p-value based weights from both SW mFAT and Improved SW mFAT for the stage presented above in Figure 1.12. The Improved SW mFAT shows the bounds placed on the p-values which work to prevent the unequal exploration and false identification of an insignificant effect size as significant. Since 00001 has a weight much closer to the other treatments, 00010 is allocated substantially more experimental units than in the SW mFAT method. As such, it gains the opportunity to collect more information which results in the p-value exiting the grey area between the bounds and being treated ultimately equal to 00001.

Figure 1.13: Visualization of Improved SW mFAT p-value weights as compared to SW mFAT. The bounds from Improved SW mFAT mitigate the randomness seen in the SW mFAT weighting.



By introducing bounds on the p-values, we are no longer treating the p-values themselves as weights, but rather using them as indicators of significance as intended. Additionally, the inclusion of the minimum p-value allows treatments with heavily significant effect sizes to continue to be explored, which will mitigate type I errors and improve our ability to identify the optimal treatment, a crucial step for following stages. This addresses the exploration inconsistencies and potential for error described in the Allocation within Stages drawback. All p-values above a certain threshold are considered equal in the weighting process, so the algorithm will not have a favored treatment among those which are likely all drawing randomly from $U(0,1)$. While it is still possible to randomly produce a low p-value, the minimum weight bound preserves the possibility for that treatment to be selected again and subsequently recover from the error.

Using this newly created weight diagonal, we calculate W as described in the previous section. These weights are then used to create a multi-nomial distribution from which to select the treatment assigned to the next experimental unit. The distribution is defined for each treatment, i,

$$\mathrm{P}(j_1 \dots j_k) = \frac{\sqrt{W_{j_1 \dots j_k}}}{\sum \sqrt{W_{j_1 \dots j_k}}}, \tag{1.23}$$

where $W_{j_1 \dots j_k}$ denotes the diagonal of W corresponding to treatment $j_1 \dots j_k$.

This process is repeated with aggregate p-values for all n experimental units in the first stage. Then, identically to SW mFAT, treatments which have significant positive effect sizes are identified and the experimenter can make a decision about the next step in the process.

These algorithm updates result in the following process being followed:

Psuedo-code: Improved SW mFAT

1: Let P denote the diagonal matrix of p-values for each effect.

2: Initialize: $b \leftarrow 1$, $P \leftarrow I$, and D_w as a diagonal matrix whose values are 1. Let d_1 be the weighted design for the weight matrix PD_w.

3: Assign the first $k+2$ experimental units one to each of the treatments and two to the control. Let P_1 be the diagonal matrix of p-values for the effects estimated under design d_1.

4: **for** b in $k+3, \ldots, N$ **do**

5: Compute d_b with the weight matrix $W = P_{b-1}D_w$. Note that d_b is an augmented design produced by adding an additional unit to the previous design, d_{b-1}.

6: Calculate the multi-nomial attribution model $\mathrm{P}(j_1 \ldots j_k)$

7: Assign the experimental unit according to the attribution model and obtain P_b.

8: If all elements of the diagonal of P_b are less than α, break for-loop and end experiment.

9: **end for**

10: Declare the best performing treatment as the optimum and stop the experiment, or update the baseline (either to an extrapolated or an observed treatment) and conduct additional SW mFAT experiments.

1.4.4 A Test to Check Extrapolation Assumptions

As discussed in Section 3.4, there is a danger inherent in using the extrapolation process without a way to check the assumptions that the interaction terms are non-negative. To minimize the risk of using the extrapolated estimate, we propose an experiment that only replicates the extrapolated treatment combination before performing a full stage with that combination set as the baseline. This produces a single mean estimate that we compare to the extrapolated estimate via a two-sample t-test. If there is a significant difference, there is reason to question the assumptions that justified the extrapolated estimate.

For example, consider $k = 3$ and the extrapolated estimate for treatment combination 111, resulting from us declaring all three BP main effects as positive. Assuming no interactions, $\mathbb{E}(\hat{\mu}_{111}) = \mathbb{E}(\hat{\theta}_{000} + \hat{\theta}_{100} + \hat{\theta}_{010} + \hat{\theta}_{001}) = \theta_{000} + \theta_{100} + \theta_{010} + \theta_{001}$. After our preliminary experiment in which we only replicate treatment 111 r_{111} times, we have the alternative estimator $\bar{Y}_{111}$ which is always unbiased for μ_{111}. Removing all assumptions about the BP interaction effects, the difference of these two estimators of μ_{111} has expected value

$$E(\bar{Y}_{111} - \hat{\mu}_{111}) = \theta_{110} + \theta_{101} + \theta_{001} + \theta_{111} \ . \tag{1.24}$$

If (1.24) is zero then we can proceed with the extrapolated baseline, as our assumptions are consistent with the data. Alternatively, if (1.24) is positive, this indicates unexpected synergies between the three factors. The experimenter can again reasonably proceed with the extrapolated baseline. However, if the result is negative, then we have evidence that there is some negative interaction term and the experimenter should perform the next stage using the best observed treatment as the baseline. These points translate to the following one-sided hypothesis:

$$H_0 : \theta_{110} + \theta_{101} + \theta_{001} + \theta_{111} \geq 0 \qquad\qquad H_1 : \theta_{110} + \theta_{101} + \theta_{001} + \theta_{111} < 0 \ , \tag{1.25}$$

with test statistic

$$T = \frac{\bar{Y}_{111} - \hat{\mu}_{111}}{s\sqrt{\frac{1}{r_{111}} + \mathrm{Var}(\hat{\mu}_{111})}} \qquad\qquad \mathrm{Var}(\hat{\mu}_{111}) = 1^T \mathrm{Var}(\hat{\theta}_A) 1 \ .$$

In this expression s is a pooled estimate of σ that includes all data collected. As we are assuming large sample sizes, the degrees-of-freedom are large enough that a normal approximation is adequate. Figures 1.14 and 1.15 demonstrate the possible outcomes and pitfalls avoided by utilizing this method.

Figure 1.14: Improved SW mFAT when extrapolation assumptions are met. "$*$" indicates the best performing treatment in that stage. Since confirmation was received that the assumptions were met, the method continues on to the extrapolated baseline, allowing for a significant reduction in experimental units required.

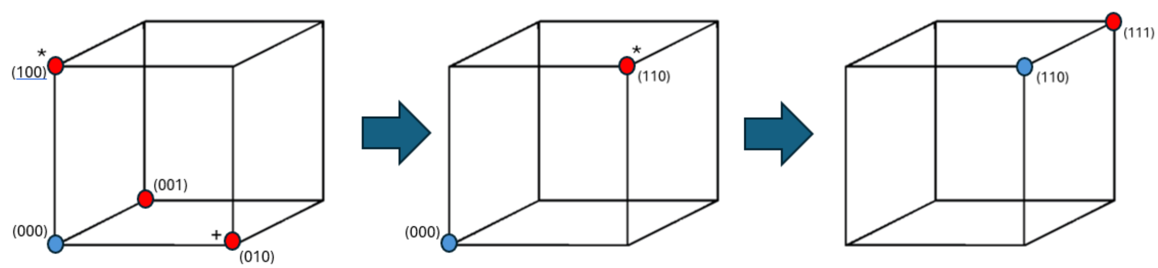
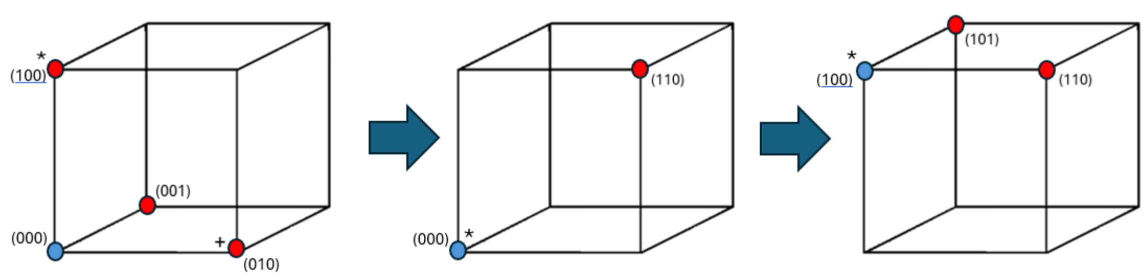


Figure 1.15: Improved SW mFAT when extrapolation assumptions are not met. "$*$" indicates the best performing treatment in that stage. By introducing the intermediate test to confirm the assumption that interactions between relevant treatments are non-negative, the method prevents research into a potentially poorly performing treatment.



While this does introduce more time and expenses, it is still less than if the experiment was run one factor change per stage. In including this step, there is more security in making the leap to the extrapolation whereas going in blindly could lead to significantly more losses than the time spent to determine if the assumptions hold true.

Psuedo-code: Extrapolation Stage 2 of Improved SW mFAT

1: Let e denote the extrapolated treatment and μ_e the expected value of that treatment.

2: Initialize $\hat{\mu}_e = \sum \hat{\theta}_i$ where $\hat{\theta}_i$ are the observed BP main effects for each factor included in the extrapolated treatment.

3: Run an experiment which replicates only the extrapolated treatment r times. 4: Calculate $\bar{Y}$, the observed mean of the extrapolated treatment from the experiment. 5: Perform a two-sample t-test comparing the $\bar{Y}$ to $\hat{\mu}_e$.

5: If the t-test returns a significantly negative result

6: Set the best observed treatment from stage one as the new baseline and repeat stage 1

7: Else

8: Set the extrapolated treatment as the baseline for the next stage of the experiment and repeat stage 1

1.5 Simulation Studies

This section presents a set of simulation studies to compare the effect size estimation and regret of the proposed heuristic to the other methods previously discussed in this paper. The following assumptions were made to most accurately imitate a true OCE. The first is strong effect heredity, which states that for any nonzero interaction effect, all the corresponding lower-order effects are also nonzero. This property is a common consequence of business practices in OCEs, as interaction effects are measured only among treatments that have positive baseline effects. The second is that k is of a sufficiently large size. This seeks to emulate the phenomenon that most organizations running OCEs run many $k = 1$ experiments at a time. In addition, it is widely accepted that in the online environment, the majority of these examined treatments do not return a significant improvement over the control. In fact, Kohavi et al. (2013) suggests a rule of thumb that only 10% - 30% of ideas demonstrate a marked improvement, based on their observations of several large online organizations. To mimic this phenomenon, most of the treatment means are assigned values less than or equal to the control's mean. These assumptions result in higher-order interaction terms frequently being not significant, as that would require all lower-order treatments to be significant in an environment where that is not common. As such, the simulations only allow treatments which change a limited number of factors to have the largest improvement over the baseline.

To meet these assumptions, treatments were obtained from factorial combinations of $k = 5$ factors, with a measurable effect size of $\delta = 0.1$, and Type I and II error rates of 0.05 and 0.2, respectively. Each of the following simulations demonstrates the behavior of the five algorithms discussed in this paper over 100 runs under the same assumptions but within different contexts. In practice, each run of the SW mFAT methods would update its total N after each stage is complete and the need for a subsequent stage is determined. All stage ones with the same assumptions would have the same N to be fully powered, as all relevant measures, δ, α, β and k are the same. However, based on the results of that stage, the number of treatments examined in stage two may vary, resulting in different N for a fully powered experiment. For example, if the first stage resulted in the new baseline being the directly observed 10000 then stage two would compare four treatments to this updated baseline. In contrast, if the new baseline is extrapolated to 11000 then there would be three treatments compared in stage two. This would obviously result in a variable max N for a fully powered process. This value may be further modified based on external constraints such as budget or prior knowledge.

For purposes of simplicity and to accommodate the MAB methods which do not have stages at which to alter their N value, the simulations all use the largest possible max N, meaning they assume 12,260 for stage one with five treatments and 10,070 for stage two with four treatments. This results in a total N of 22,330. For $k=5$ and $k=4$ respectively, these values are a conservative choice to guarantee a power of 80% for detecting any main effects of at least size $\delta=0.1$ with $\alpha=0.05/k$. In practice, this decision would produce a slightly more powered experiment which allows the methods to accurately detect effect sizes slightly smaller than 0.1. In very specific circumstances, this could allow for unexpected early termination if treatments all have effect sizes larger than this value, even if some were less than the originally determined δ. The treatment mean values selected in this simulation study are designed to prevent this from occurring. The treatment means used for the simulations are listed in Table 1.7 with updates unique to the simulation study noted at the beginning of each subsection.

Table 1.7: Simulation Setting 1 - treatment means for $k=5$

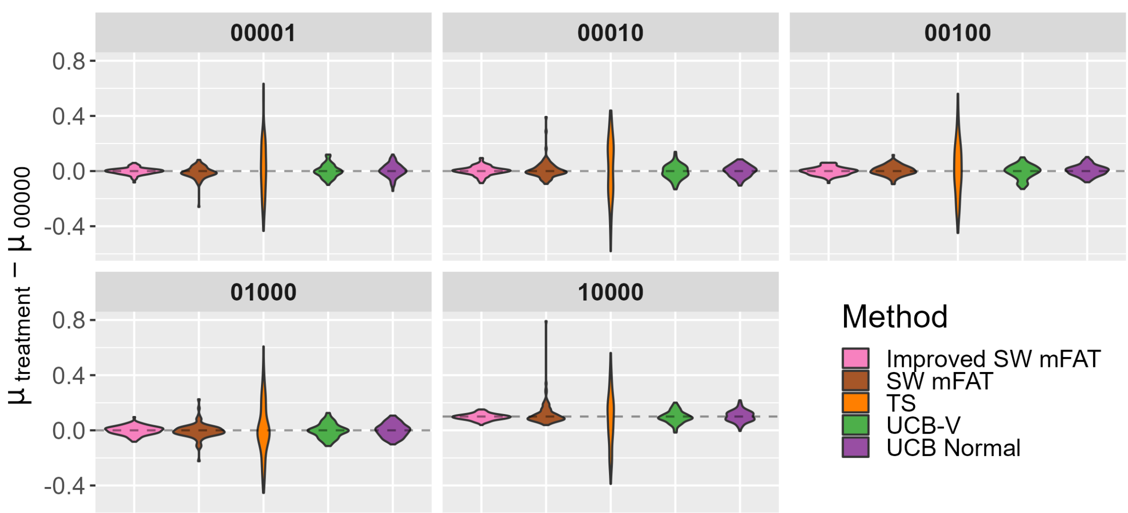
Treatment	**00000**	**10000**	**01000**	**00100**	**00010**	**00001**	**11000**	**10100**
μ	3	3.1	3	3	3	3	3	3
	10010	**10001**	**01100**	**01010**	**01001**	**00110**	**00101**	**00011**
	3	3	3	3	1	3	1	1
	11100	**11010**	**11001**	**10110**	**10101**	**10011**	**01110**	**01011**
	3	3	3	3	1	1	1	1
	01101	**00111**	**11110**	**11101**	**11011**	**10111**	**01111**	**11111**
	1	3	1	1	1	1	3	3

The practical success of a OCE is measured by the correct identification of the optimal treatment at the lowest possible cost. Since these simulations are designed to run to completion with no early stopping, the differences in price (or experimental units) are briefly touched upon but not addressed in depth. Instead, focus is placed on accuracy of effect size, regret and identification of the optimal treatment.

1.5.1 One Significant Main Effect

In this section, the simulation presents the situation in which only one main effect, and thus no higher-order effects, is significant. Main effect estimates under Improved SW mFAT, SW mFAT, Thompson Sampling, UCB-N and UCB-V can be found in Figure 1.16.

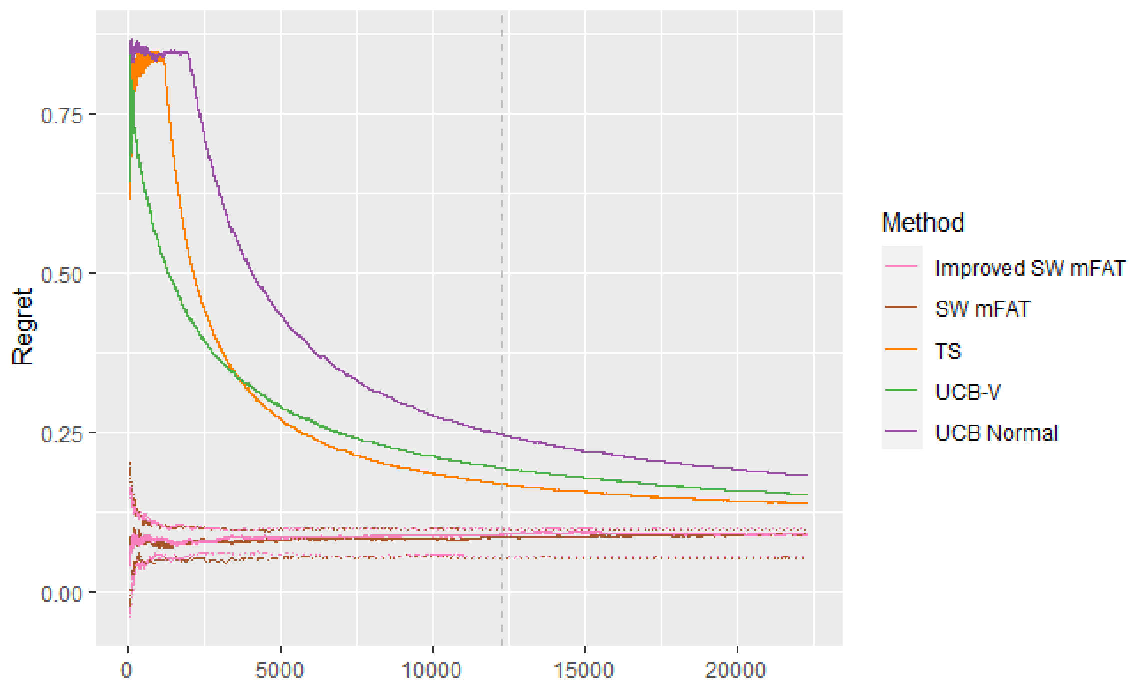
Figure 1.16: Distribution of $\mu_{treatment} - \mu_{00000}$ under Setting 1.



As all methods return unbiased estimates, we note that all main effects are centered around the expected effect size. Thompson sampling resulted in the highest variability, which is expected as its design prioritizes minimization of regret over all other results. The UCB algorithms performed similarly to one another, both resulting in a similar variability to the Improved SW mFAT method. The original SW mFAT method shows a similar variability shape to the other three with the exception of the long tails. These tails result from a combination of a higher type I error rate than the other methods and the lack of recovery for treatments with non-significant effect sizes being miscategorized as significant. Having addressed these issues, the Improved SW mFAT method displays substantially smaller tails than SW mFAT.

Regret curves for Improved SW mFAT, SW mFAT, Thompson Sampling, UCB-N and UCB-V can be found in Figure 1.17 for the same 100 replicates.

Figure 1.17: Regret plot under Setting 1.



The median regret curves for all five methods are plotted as the solid lines alongside the 25^{th} and 75^{th} quantiles for the SW mFAT methods as the dashed lines. These quantiles were excluded from the MAB methods for these graphs for ease of understanding. For an example with the quantile lines see Figure 1.3. The regret for the MAB algorithms will always be high at the start of an experiment, as they explore all possible treatment combinations in order to tune their process towards the optimal treatment. In contrast, by focusing their efforts on only main effects from the beginning, the SW mFAT designs minimize their initial regret and maintain that lowered regret throughout the simulation. It is worth noting that this minimized regret is a result of the design decision to focus on main effects, but is not the goal of these methods. By tailoring the heuristic to specifically identify small incremental changes within the treatments, regret is simply inherently limited.

Both SW mFAT methods center very closely to 0.1 as their regret for the duration. This results from them quickly identifying the largest effect size, belonging to treatment 10000, and thus not replicating it frequently. Instead, focus is placed on exploring the other treatment combinations, all of which have expected effect sizes of 0. When compared to the best treatment's effect size of 0.1, regret is approximately 0.1. The lower quantile line is caused by the runs which, for that experimental unit chose to allocate it to treatment 10000.

The vertical grey line indicates when the experiment would be expected to stop for the SW mFAT methods. In this case, the expected stopping point is the end of stage one, as with only one treatment returning a significant effect size, the strict heredity assumption dictates that there will not be an improvement found in later stages. This would prevent the use of an additional 10,070 experimental units. As such, the most meaningful comparisons will occur prior to this line.

However, even with the inclusion of the buffer experimental units, the Improved SW mFAT method slightly outperforms the MAB methods and SW mFAT in correctly identifying the optimal treatment. Table 1.8 shows the distribution of treatments identified as the optimal by the five methods.

Table 1.8: Number of times each treatment was identified as the optimal treatments by each of the five examined methods across the 100 runs in Setting 1.

	Improved SW mFAT	SW mFAT	Thompson Sampling	UCB-N	UCB-V
00000	0	0	29	3	9
10000	100	93	23	97	91
01000	0	3	7	0	0
00100	0	1	12	0	0
00010	0	3	16	0	0
00001	0	0	13	0	0
Other	0	0	0	0	0

1.5.2 Multiple Significant Main Effects with No Interaction

The next simulation considers performance when there is a significantly positive 2FAT treatment and the assumption that the interaction between the two comprising treatments is negligible holds. The updated treatment means are listed in Table 1.9.

Table 1.9: Treatment means for $k = 5$ - Simulation Setting 2

Treatment	**00000**	**10000**	**01000**	**00100**	**00010**	**00001**	**11000**	**10100**
μ	3	3.2	3.1	3	3	3	3.3	3

As expected, Figure 1.18 shows all five methods' estimated main effects distribution is similar to the previous simulation. SW mFAT shows a more distinct positive tail for 10000 and 01000, resulting from the lack of replication of the two treatments with significant effects.

Figure 1.19, the effect size estimates for the relevant two factor treatment combinations, are of more interest. SW mFAT and Improved SW mFAT both extrapolated in stage two so we examine the effect size estimates for the three treatments under consideration as well as 11000 the ’new control’.

Figure 1.18: Distribution of $\mu_{treatment} - \mu_{00000}$ for treatments in Stage One under Setting 2.

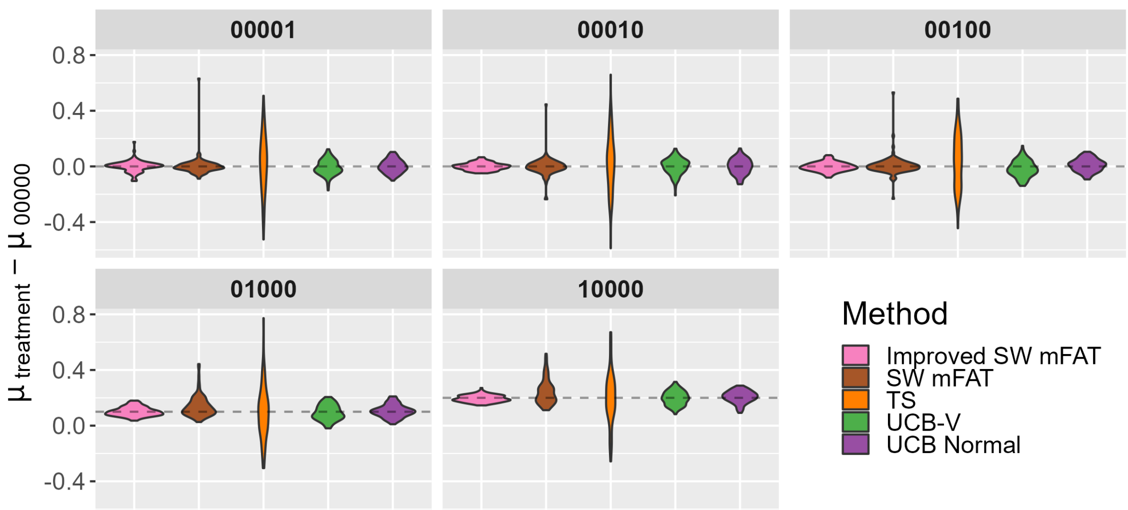
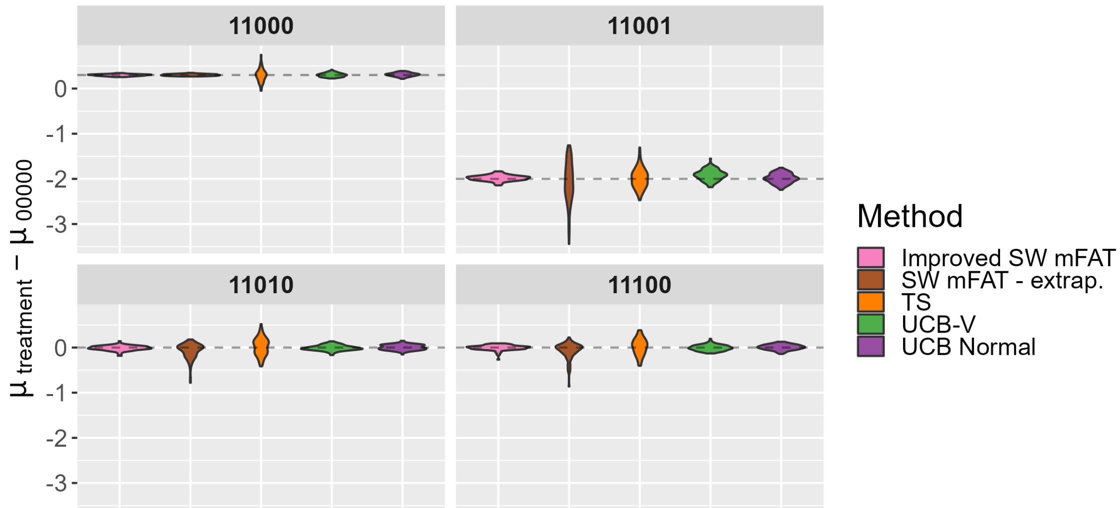


Figure 1.19: Distribution of $\mu_{treatment} - \mu_{00000}$ for treatments in Stage Two under Setting 2.



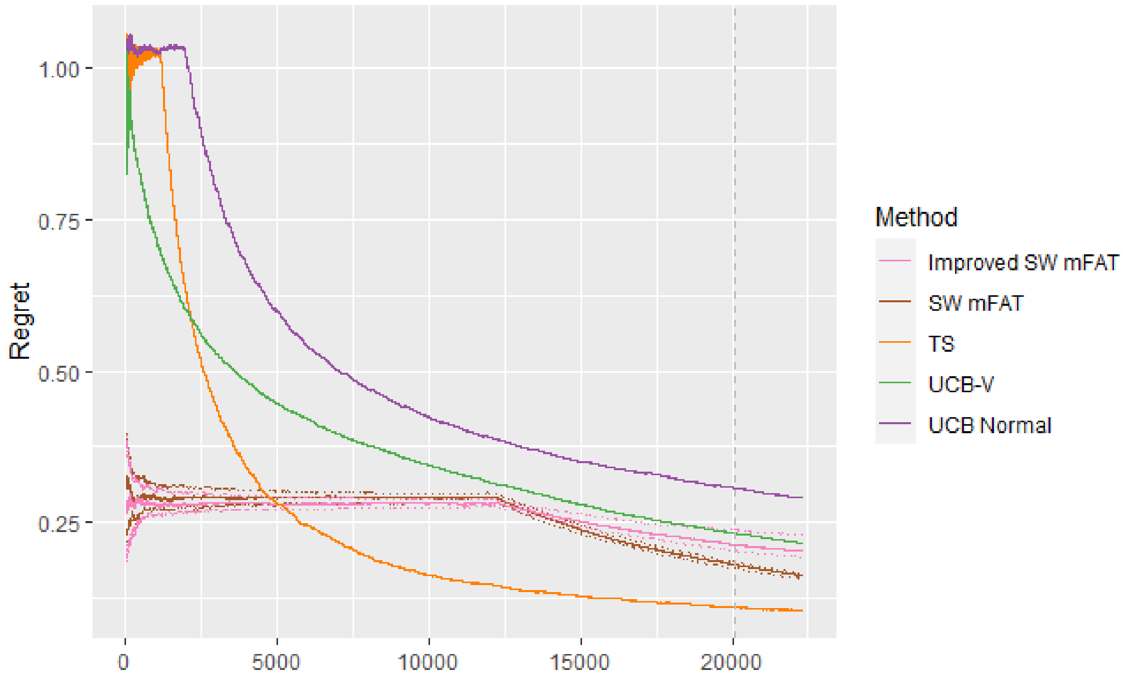
Performance varied among the two factor treatment combinations. In particular, the graph for the estimate of treatment 10001, with $\mu = 1$, shows a wider variability across the board. This is because all the methods allocated fewer experimental units to this treatment due to its clear significantly different mean. The three MAB algorithms stopped replicating it quickly because it demonstrated worse performance. In contrast, SW mFAT and Improved SW mFAT slow on replicating it because the difference from the control became significant, though no care was taken for whether the effect size was positive or negative. Although small, it is clear that the variance for Improved SW mFAT is slightly smaller than SW mFAT due to the increased replication of treatments with significant effect sizes.

The SW mFAT methods performed similarly to the UCB methods in the other treatment combinations as well. The UCB methods focused on 11000 to exploit as the clear optimal treatment combination. This led to lower replication and thus higher variability in the other two treatments, though not as low as 10001. SW mFAT methods purposefully allocated relatively equal experimental units in an effort to explore these other two treatment combinations after identifying 11000 as significant. As previously observed, SW mFAT continued to have longer tails indicating type I errors.

The regret functions for this setting are shown in Figure 1.20. The MAB algorithms' curves look very similar to those from the previous simulation as their process did not change despite the treatment means altering. For the SW mFAT methods, regret is still minimized in the beginning by the SW mFAT methods due to the focus on main effects, however the optimal treatment is not identified until stage two, which is when the dip in regret begins for both methods.

The regret functions for Thompson Sampling, UCB-V and both SW mFAT methods now show very similar regret at the end of the experiment. This is a result of the larger effect size of the optimal treatment. This makes it easier to identify that treatment for the MAB algorithms and thus exploit it to lower regret faster. Instead of optimizing regret, the SW mFAT methods are focused on exploration. Thus the larger the largest main effect is, the higher the regret curve. This exemplifies why the SW mFAT methods are optimized for the environment that most OCEs find themselves in; attempting to identify very small improvements over the control.

Figure 1.20: Regret plot under Setting 2.



The vertical grey line again indicates when the SW mFAT methods would have been expected to stop, this time at the end of a stage two with only three treatment combinations being examined. With a conservative N estimate of 7,855 for this stage, that would result in the experiment ending 2,215 experimental units earlier. However, because Improved SW mFAT included the intermediate test to confirm the assumption, it allocated an additional 3,142 experimental units over the SW mFAT. As such, when extrapolating the Improved SW mFAT does require extra experimental units in order to prevent unsafe exploration.

Similar to the previous simulation, Table 1.10 shows that the distributions of correct identifications of the optimal treatments are essentially equal between the five methods, with Improved SW mFAT, UCB-N and UCB-V all identifying the optimal treatment every time.

Table 1.10: Number of times each treatment was identified as the optimal treatments by each of the five examined methods across the 100 runs in Setting 2.

	Improved SW mFAT	SW mFAT	Thompson Sampling	UCB-N	UCB-V
00000	0	0	9	0	0
11000	100	75	56	100	100
10000	0	20	18	0	0
01000	0	3	4	0	0
00100	0	1	1	0	0
00010	0	1	3	0	0
00001	0	0	4	0	0
10100	0	0	2	0	0
10010	0	0	3	0	0

1.5.3 Multiple Main Effects with Negative Interaction

This simulation considers performance when there is a significant higher-order effect but the interaction between the two significantly positive one factor treatments is negative. The treatment means are listed in Table 1.11.

Table 1.11: Treatment means for $k = 5$ - Simulation Setting 3

Treatment	**00000**	**10000**	**01000**	**00100**	**00010**	**00001**	**11000**	**10100**
μ	3	3.3	3.1	3	3	3	3.2	3

As demonstrated in Figures 1.21 and 1.22, the one and two factor treatments have similar estimate centers and variability around the true treatment means as in previous simulations for four of the five methods. The SW mFAT model is not displayed in these images as there were only a handful of simulations which resulted in the testing of these treatment combinations. This is a result of the extrapolation assumptions in that algorithm. While μ_{11000} is an improvement over μ_{00000}, it is not the optimal final treatment selection. This demonstrates the impact of violating the assumption of no interactions between treatments.

Figure 1.21: Distribution of $\mu_{treatment} - \mu_{00000}$ for treatments in Stage One under Setting 3.

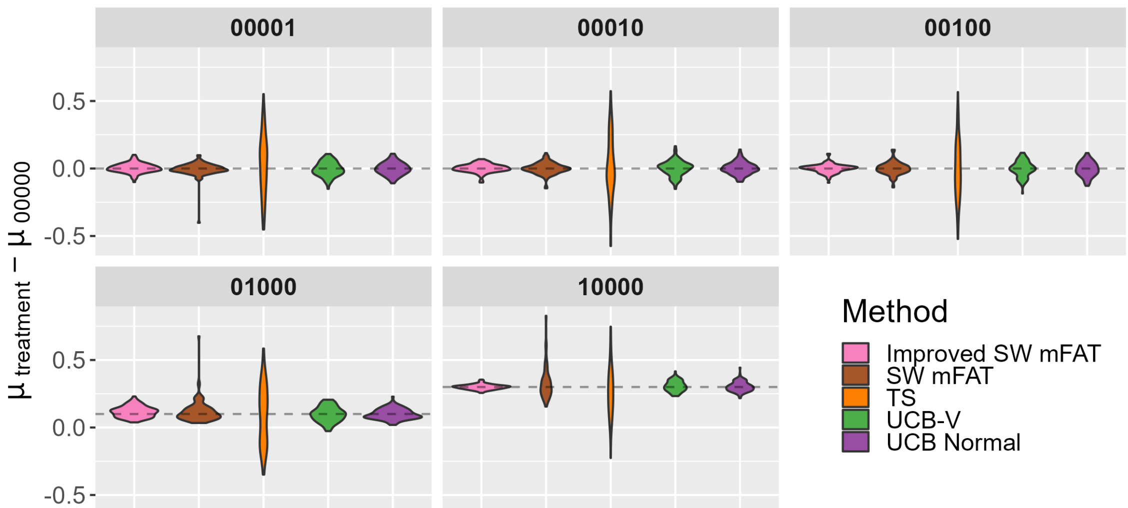
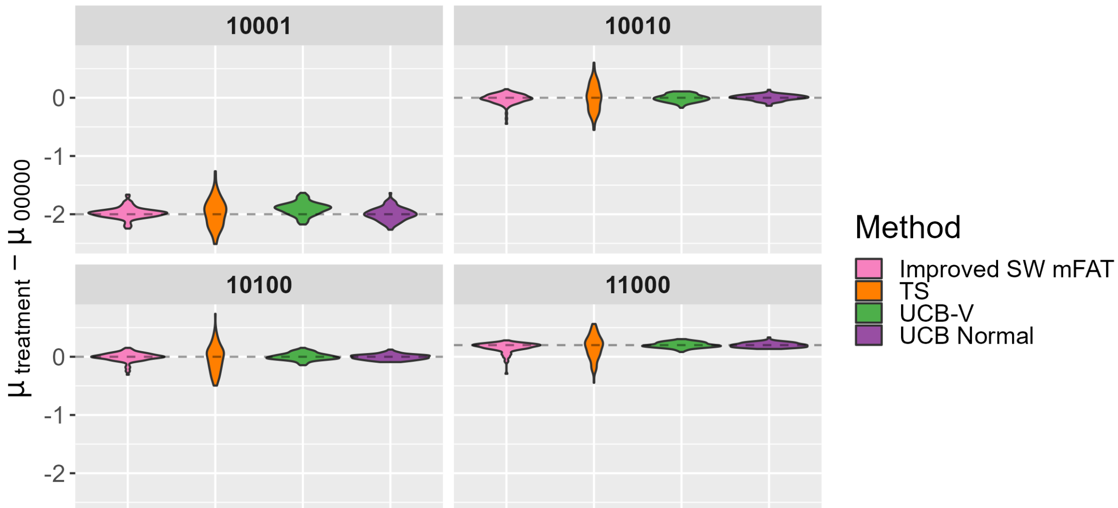


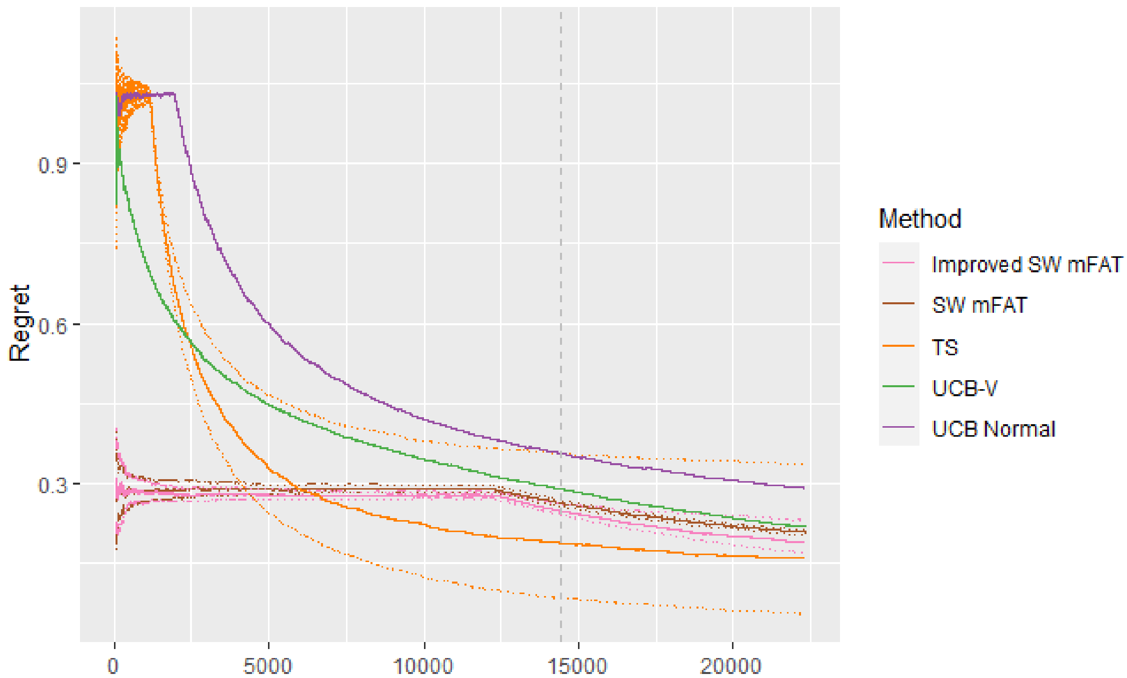
Figure 1.22: Distribution of $\mu_{treatment} - \mu_{00000}$ for treatments in Stage Two under Setting 3.



This discrepancy does not appear in the regret curves in Figure 1.23. The MAB algorithms continued to show consistency across simulation settings. The Improved SW mFAT regret now emulates the logic of the first simulation in stage one but at a higher initial regret. Treatments 10000 and 01000 are both quickly identified as significant and so receive lower allocation. The other treatments are explored more, which results in the regret hovering around 0.3 as that is the best treatment's effect size. Upon the stage ending and identifying two significant effect sizes, the intermediate test for interactions was completed and failed. Thus, like the first simulation, 10000 was set as the baseline for stage two. All the other treatments are identified as having a significant effect size as compared to 10000 and the stage terminates early. Remaining replicates are allocated to the identified optimal treatment for the sake of graphing.

SW mFAT performs the same way for stage one. However, it immediately chooses to extrapolate for stage two. This results in a stage two baseline of 11000 with $\mu = 3.2$, which when compared to the three relevant treatment combinations all with $\mu = 3$, also results in an early termination. These early terminations lead to a similar regret curve as both are assigning the majority of the experimental units for the stage to the optimal treatment.

Figure 1.23: Regret plot under Setting 3.



Improved SW mFAT would allocate N of 22,330, plus the 3,142 for the interim stage test, though the early stopping will lower this value depending on how long the algorithm takes to recognize all treatments as significantly different. The SW mFAT would allocate $N = 20,115$, though again early stopping would limit this. The grey vertical line on the regret chart is placed at the maximum allocations for SW mFAT as the graph ends at the maximum allocation for Improved SW mFAT.

The real consequence of this break in decision making occurs in the distribution of optimal treatments determination as seen in Table 1.12. While Improved SW mFAT remains in close contention with the MAB algorithms, the SW mFAT is mostly incorrect in its assertions. This is the situation in which the protection to the company that comes at the cost of briefly waiting to test the intended extrapolated baseline pays off.

Table 1.12: Number of times each treatment was identified as the optimal treatment by each of the five examined methods across the 100 runs in Setting 1.

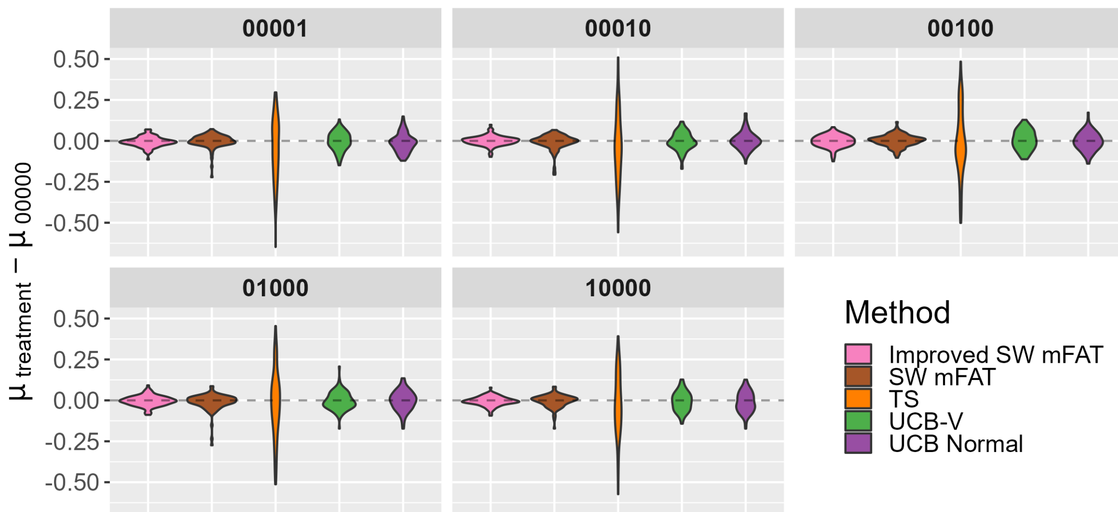
	Improved SW mFAT	SW mFAT	Thompson Sampling	UCB-N	UCB-V
00000	0	0	20	0	0
10000	100	0	47	100	100
01000	0	0	4	0	0
00100	0	0	2	0	0
00010	0	0	2	0	0
00001	0	0	6	0	0
11000	0	100	17	0	0
10100	0	0	2	0	0

1.5.4 All Treatments have Insignificant Main Effects

The final simulation considers the situation in which there are no treatments significantly better than the control. In this case, all treatment means are set to 3.

This simulation is unique in that there is no regret from any of the models, since all treatments produce equal effect sizes of 0. As such, a regret plot is not offered for this section. However, in an ideal situation, the control would be selected as the final model for this simulation, as none of the other treatments show an improvement over it and it is most cost-effective from the experimenter's perspective to continue using the status quo. In this case, the positive tails in the distributions of main effect estimates in Figure 1.24, which match the trends we've seen in previous simulations, are the times the result would have the experimenter make an ill-advised decision.

Figure 1.24: Distribution of $\mu_{treatment} - \mu_{00000}$ when there are no treatments significantly different from the control.



Improved SW mFAT and the two UCB methods continue to provide the least variance around the main effect estimates, while SW mFAT has a limited number of type one errors. Thompson sampling again demonstrates its sacrifice of precision in estimation in order to minimize regret. Note also the similarity in variability within each method across treatments. For MAB methods, this is a result of being unable to identify a clear optimal treatment, so exploration happens on all equally. Both SW mFAT and Improved SW mFAT instead are focused on exploring these treatments because they are not showing significant differences, providing the experimental units needed to discover a smaller, but significant, effect size, if present.

The MAB algorithms distributed the selection of optimal treatments across all treatments combinations, although it was weighted towards the control. This could have potentially heavy impacts on users as they would choose to change aspects of their object of interest based on what is actually an arbitrary selection. Alternatively, the Improved SW mFAT guards against this exact misconception and always selects the control should no significant findings otherwise be presented.

This edge case exemplifies the importance of accurate estimates, as returning no significant findings is plausible for an OCE. Correctly identifying that there is no distinct difference between any of the treatments of interest and the control could save users and industries a lot of time and money.

1.6 Discussion

This paper improves the efficacy of the SW mFAT method presented in Larsen (2023) by addressing some methodological weaknesses, allowing it to apply to problems with fewer restrictions than those discussed previously. These contributions are:

1. **Minimizing Potential Type I Error** We addressed type I error in two ways. By redefining the initial design to have α as a family-wise error rate instead of pairwise, we mitigated the phenomenon of having a higher error rate at large values of k. Further, by implementing the minimum p-value bound during the weighting calculations for allocation, we encouraged further exploration into treatments which had been found significant. By doing so, treatment combinations which were not in truth significantly different from the control might recover and be explored further later in the stage.

2. **Correct Replication of the Control** The over-replication of the control in each stage was due to the assumption that we needed to accurately estimate the mean of the control. This assumption does not make sense in the context of OCEs nor is it required for the algorithm, which focuses purely on measuring effect size. Therefore, we can

remove the allocation assigned to the control purely to fine-tune its estimate. By limiting replication of the control to only that required to optimize the estimates of effect size, the heuristic could utilize more experimental units in exploring treatments not yet found to be significantly different from the control.

3. **New Allocation Method** We incorporated the small changes noted above into an entirely new allocation method of the n experimental units in a stage. By eliminating the use of blocks for allocation purposes, we needed an allocation method which could support individual experimental unit allocation. In summary, this process calculated the p-values of each treatment, bounded them to prevent mistaken assertions of both strong and weak significance, and created a multi-nomial distribution with which to assign the experimental unit to a treatment combination. By doing so, we addressed the issue of calculating block sizing, maintained equal exploration of treatment combinations resulting in non-significantly different effect sizes, and maintained the improvements identified above.

4. **Strategy for Safe Extrapolation** We introduced a test between stages in order to confirm that the assumption of interaction effects being non-negative held if extrapolation is desired. By implementing this test, we prevented the method from failing to identify the treatment with the largest effect size simply because it skipped testing that treatment all together in scenarios with negative interaction effects.

One potential extension to the current method is examining the situation in which the first stage contains two factor treatment combinations. This is possible in the current design, however it would necessitate the re-introduction of the minimum variance weighting to the allocation process as the minimum variances would vary between all possible treatments in the stage. Another point of research revolves around the handling of negative significant effect sizes. While Improved SW mFAT would encourage continued replication of this treatment in an effort to allow recovery if the significance result is incorrect, it also leaves the door open to replicating a truly negative treatment. This might be addressed by creating different thresholds for the p-value weightings dependent on effect size value or redefining the function bounding the p-values to be continuous instead of simply using cutoffs.

Other extensions come to light when looking to other overlapping questions in recent research. While SW mFAT effectively accounts for erroneous false-positives of the effect estimates, this heuristic has yet to take advantage of the body of literature surrounding optional stopping. Given the sequential nature of the proposed heuristic, it seems only natural to incorporate sequential p-values into the framework. Another potential approach to early termination which maintains the current p-value definition can be found in the process of bounding the p-values. These bounds are essentially saying that p-values between α and the selected upper quantile value correspond to effects that lay in a gray area of significance. This may imply that p-values larger than the quantile correspond to effects we can discard with a level of confidence as being unimportant. If that is the case, then early termination guidelines could be implemented when all treatments result in either compelling significant or insignificant effect sizes. These guidelines are another point of potential study.

In addition to sequential stopping, further opportunities present themselves in developing ways to handle type I error. The introduction of a less conservative error adjustment or calculating the weighting of p-values over a sliding window are two potential alternatives. Additionally, recent developments in the OCE community have highlighted an interest in hold-out experimentation. Implementing baseline parametrization in the online hold-out designs outlined by Shao and Burke (2022) may prove beneficial for assessing the cumulative impact of several successful variants in comparison to the sum of each variant's individual effects.

REFERENCES

Agrawal, R. (1995). Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078.

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings.

Allen-Moyer, K. and Stallrich, J. (2022). Incorporating minimum variances into weighted optimality criteria. *The American Statistician*, 76(3):262–269.

Amadio, B. (2020). Multi-armed bandits and the stitch fix experimentation platform. `https://multithreaded.stitchfix.com/blog/2020/08/05/bandits/`. (Accessed on 03/21/2022).

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256.

Chen, A., Sun, C.-Y., and Tang, B. (2021). Selecting baseline designs using a minimum aberration criterion when some two-factor interactions are important. *Statistical Theory and Related Fields*, 5(2):95–101.

Haizler, T. and Steinberg, D. M. (2021). Factorial designs for online experiments. *Technometrics*, 63(1):1–12.

Karunanayaka, R. C. and Tang, B. (2017). Compromise designs under baseline parameterization. *Journal of Statistical Planning and Inference*, 190:32–38.

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

Lai, T. L., Robbins, H., et al. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Larsen, N. A. (2023). *Advances in the Design of Online Controlled Experiments with Network and Baseline Factorial Effects.* PhD thesis, North Carolina State University.

Li, P., Miller, A., and Tang, B. (2014). Algorithmic search for baseline minimum aberration designs. *Journal of Statistical Planning and Inference*, 149:172–182.

Li, W., Liu, M.-Q., and Tang, B. (2022). A systematic construction of compromise designs under baseline parameterization. *Journal of Statistical Planning and Inference*, 219:33–42.

Mak, S., Zhao, Y., Hoang, L., and Wu, C. J. (2022). Tsec: a framework for online experimentation under experimental constraints. *Technometrics*, 64(4):513–523.

Miller, A. and Tang, B. (2016). Using regular fractions of two-level designs to find baseline designs. *Statistica Sinica*, pages 745–759.

Mukerjee, R. and Huda, S. (2016). Approximate theory-aided robust efficient factorial fractions under baseline parametrization. *Annals of the Institute of Statistical Mathematics*, 68(4):787–803.

Mukerjee, R. and Tang, B. (2012). Optimal fractions of two-level factorials under a baseline parameterization. *Biometrika*, 99(1):71–84.

Mukerjee, R. and Tang, B. (2016). Optimal two-level regular designs under baseline parametrization via cosets and minimum moment aberration. *Statistica Sinica*, pages 1001–1019.

Optimizely (2022). Multivariate testing vs a/b testing - optimizely. `https://www.optimizely.com/optimization-glossary/multivariate-test-vs-ab-test/`. (Accessed on 06/20/2022).

Pukelsheim, F. (2006). *Optimal design of experiments*. SIAM.

Scott, S. L. (2010). A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658.

Scott, S. L. (2015). Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry*, 31(1):37–45.

Shao, S. and Burke, C. (2022). Etsy engineering - understanding the collective impact of experiments. `https://www.etsy.com/codeascraft/understanding-the-collective-impact-of-experiments`. (Accessed on 03/08/2023).

Siroker, D. and Koomen, P. (2013). *A/B testing : the most powerful way to turn clicks into customers*. Hoboken : Wiley, 2013.

Stallings, J. W. and Morgan, J. (2015). General weighted optimality of designed experiments. *Biometrika*, 102(4):925–935.

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

Thompson, W. R. (1935). On the theory of apportionment. *American Journal of Mathematics*, 57(2):450–456. Publisher: The Johns Hopkins University Press.

Wu, C. J. and Hamada, M. S. (2011). *Experiments: planning, analysis, and optimization*, volume 552. John Wiley & Sons.

Yan, Z. and Zhao, S. (2023). Optimal fractions of three-level factorials under a baseline parameterization. *Statistics & Probability Letters*, 202:109902.

Yan, Z. and Zhao, S. (2025). Optimal s-level fractional factorial designs under baseline parameterization. *Journal of Statistical Planning and Inference*, 236:106242.