# Adaptive Treatment Assignment in Experiments for Policy Choice

(preliminary draft)

Maximilian Kasy* Anja Sautmann†

April 15, 2019

**Abstract**

The goal of many experiments is to inform the choice between different policies. However, standard experimental designs are geared toward point estimation and hypothesis testing. We consider the problem of treatment assignment in an experiment with several cross-sectional waves where the goal is to choose among a set of possible policies (treatments) for large-scale implementation. We show that optimal experimental designs learn from earlier waves by assigning more experimental units to the better-performing treatments in later waves. We discuss a computationally tractable approximation of the optimal design, based on a modification of Thompson sampling. Calibrated simulations and theoretical results demonstrate improvements in welfare, relative to conventional designs as well as standard Thompson sampling. Our setting is related to but different from multi-armed bandit settings. The focus on the highest-performing policies is not driven by an "exploitation" motive, but by optimal learning about the best policy choice.

Keywords: Experimental design, field experiments, optimal policy
JEL Codes: C93, C11, O22

*Department of Economics, Harvard University, maximiliankasy@fas.harvard.edu
†Massachusetts Institute of Technology, sautmann@mit.edu.
We thank the seminar participants at the development economics retreat at Harvard in winter 2018, at CMU, and Syracuse. All errors and opinions are our own.

# 1 Introduction

One of the biggest methodological shifts in public policy research, and the social sciences more broadly, has been the introduction of randomized control trials (RCTs) as a tool. RCTs have been particularly influential in development economics (Banerjee et al., 2016). The main objective of an academic researcher conducting an RCT is typically to generate a point estimate and standard errors of the treatment effect, in order to test the null hypothesis that there is no average effect. The research design is chosen to maximize power for tests of this null, for example by assigning an equal number of units to different treatments, and by stratifying the sample by pre-determined covariates (see for instance Athey and Imbens 2017). Such RCTs are designed to answer the question "Does this program have a significant effect?"

However, the objective of an NGO or government who considers conducting an experiment to evaluate its programs is often slightly different: instead of estimating effect size, they are interested in identifying and implementing the best out of several possible policies or policy variants. In other words, they would like to answer the question, "which program will have the largest effect"? We will show that the objective of informing policy choice leads to design recommendations that are qualitatively different from standard RCT recommendations.

Take a small NGO with the goal of improving newborn health by encouraging "kangaroo care" (skin-to-skin care) for prematurely born babies. Kangaroo care, implemented correctly and used consistently, works in principle (Conde-Agudelo et al., 2012). However, multiple implementation choices determine uptake, from setting the right incentives for health-care providers to educating mothers. How should the NGO make these choices? A conventional RCT is not optimally suited to the problem, not least because the number of program variants to test may rapidly outgrow feasible sample sizes.

We argue that the NGO should run an experiment in multiple waves. Initially, they should try many different variants. As they learn which options perform better, they should focus the experiment on these highest-performing options. This allows them to identify the best option for implementation at scale faster and with greater precision. As this example suggests, one could think about our proposal as a principled approach for running pilot studies, or "tinkering" with policy options, in the spirit of "the economist as plumber" (Duflo, 2017). We will show that our proposal leads to consistently better policy recommendations than a standard RCT, or conversely that

smaller samples are needed to achieve the same expected outcomes as with a standard experimental design.

We consider an experimental setting with multiple waves of experimental units, and multiple treatments (policies). We assume that the outcome of interest is binary. At the beginning of each wave, the number of units assigned to each treatment arm is decided. After conclusion of the wave, prior beliefs about treatment effects are updated based on the observed success rates (outcomes) in the different treatment arms. Then treatments are assigned for the next wave, based on these updated beliefs. Once the experiment is concluded, one of the treatments is picked for full-scale implementation. The objective is to maximize the average outcomes for this full-scale implementation, net of the costs of treatment.

The setting is closely related to the well-known "multi-armed bandit" problem (cf. Bubeck and Cesa-Bianchi, 2012; Weber et al., 1992), but with the key difference that there is no "exploitation" motive, and thus no exploitation-exploration tradeoff. This is because in our setting the goal is to maximize outcomes after the experiment is concluded, but not during the experiment. There are several situations where this is justified. First, ongoing testing may not be feasible for various reasons, so that it is best to optimize the experiment for speedy learning. It may simply not be practical in the long run to implement treatment arms in parallel, such as asking nurses to follow different protocols for kangaroo care for different patients; the costs of surveying, data collection, and implementation of multiple treatment arms may be prohibitive; or there may be resistance from politicians, from the public, or from implementers to conduct a "never-ending experiment." This is the case even for many online experiments of internet companies, which are run for a limited duration only, since continuing experimentation would require monitoring by qualified employees.

Second, in some applications the outcomes during the experiment are not relevant. For example, consider a researcher who is piloting different survey protocols to maximize response rates: since the effective sample composition changes during the pilot, the pilot observations cannot be used in the final data analysis. Another example are practice runs for a future event, such as testing emergency evacuation protocols: here, all outcomes are realized when the emergency actually occurs, whereas prior tests of the protocol have no welfare consequences.

The policy choice problem described above defines a finite horizon dynamic stochastic optimization problem. The actions in each wave (period) are the different possible

treatment assignments; the state are current beliefs over treatment effects; and transitions between states from period to period are determined by experimental outcomes. In the final period, the action consists in the choice of policy for implementation, after which welfare is realized as the average per-unit outcome net of costs.

This optimization problem can in principle be solved analytically using backward induction, and we discuss analytic solutions and their qualitative features for some small-scale examples below. In more realistic settings, however, finding exact solutions quickly becomes infeasible, due to exploding state and action spaces. These constraints motivate the use of approximate solutions that are computationally feasible. In order to overcome the computational constraint, we propose the following assignment algorithm, which is a modified version of so-called Thompson sampling (Russo et al., 2018). Thompson sampling was originally proposed as a solution to multi-armed bandit problems. In Thompson sampling, each unit is assigned to a given treatment $d$ with probability equal to the posterior probability $p_t^d$ (given past outcomes) that it is in fact the optimal treatment. This prescription is easy to implement, by sampling just one draw of the parameter vector from the posterior, and picking the optimal treatment corresponding to this parameter vector.

We modify this prescription in two ways. First, rather than independently assigning each unit to a treatment based on the probabilities $p_t^d$, we assign a corresponding share of each wave to the different treatments. Second, and more importantly, we replace the assignment shares $p_t^d$ by shares equal to $q_t^d = S_t \cdot p_t^d \cdot (1 - p_t^d)$, where $S_t$ is a normalizing constant. The shares $q_t^d$ would arise if we ran conventional Thompson sampling sequentially, within a given wave, but forced the algorithm to never assign the same treatment twice in a row.

We show, using both simulations and theoretical results, that this modification improves expected welfare. It avoids assigning more than 50% of the sample to the highest-performing treatment, and in large samples it equalizes power for rejecting each of the sub-optimal treatments. This behavior is optimal for the convergence rate of welfare, while standard Thompson sampling is not, as discussed in Section 5.

As an extension to the baseline model, we consider the case where some observable covariates can be used to target treatment assignment. In this setting, the optimal treatment assignment algorithm during the experiment also conditions on the strata formed by these covariates. We propose to use a hierarchical Bayes approach in order to form posterior beliefs about the treatment success rate of each treatment arm within

each stratum. This approach allows one to learn about effect heterogeneity and optimal targeting, while at the same time importing information across strata in an optimal way.

We provide extensive simulation evidence on the performance of modified Thompson sampling compared with alternative assignment algorithms, in particular a non-adaptive RCT (with equal treatment arm size) and original Thompson sampling.We evaluate these algorithms according to the loss that is incurred from picking another than the highest-performing treatment option with some probability. Our simulations use parameters and sample sizes calibrated to data from three published experiments in development economics (Ashraf et al., 2010; Bryan et al., 2014; Cohen et al., 2015).

Two important patterns are confirmed by these simulations. First, modified Thompson sampling consistently performs better than standard Thompson sampling, which in turn outperforms conventional non-adaptive designs: the distribution of welfare (across simulations) under modified Thompson sampling stochastically dominates the alternatives, and as a result, modified Thompson sampling generates higher average welfare and a higher probability of picking the optimal treatment. Second, as would be expected, the gains from adaptive sampling designs are larger when the experiment is divided into more waves, for the same total sample size. The same patterns hold when considering assignment stratified on covariates, and corresponding targeted treatment assignment policies.

We demonstrate the feasibility of our proposal in an experiment that uses the roll-out of a phone information campaign to rice farmers in Odisha to test different enrollment protocols. [THIS IS WORK IN PROGRESS.]

The idea of adaptive sampling is almost as old as the idea of randomized experiments; see for instance Thompson (1933). Adaptive experimental sampling designs have been used in clinical trials (Berry, 2006), and in the targeting of online advertisements (Russo et al., 2018), but they have not yet entered the standard toolkit for RCTs in economics, see e.g. Duflo and Banerjee (2017).

A large theoretical and practical literature addresses adaptive designs, focused in particular on the multi-armed bandit problem. We discuss the differences between our setting and the bandit problem in Section 2.4. Under some conditions (separability across treatment arms, infinite horizon), the optimal solution to the Bandit problem can be expressed in terms of choosing the arm corresponding to highest "Gittins index," cf. Weber et al. (1992). In practice, most applications use heuristic algorithms rather than

solving for the optimal assignment. Two popular algorithms are the Upper Confidence Bound algorithm (UCB), and Thompson sampling (Russo et al. (2018)). A fairly recent literature characterizes the expected regret of these algorithms, see for example Bubeck and Cesa-Bianchi (2012). Generalizations of the Bandit problem are discussed under the name of reinforcement learning in the machine learning literature (Ghavamzadeh et al. (2015) and Sutton and Barto (2018)). Lastly, Russo (2016) considers a problem closely related to ours, namely the problem of maximizing the probability of picking the best treatment (rather than maximizing expected welfare). Our theoretical analysis in Section 5 below will draw on insights from this paper.

In the next section, we introduce our formal setup, solve for optimal treatment assignments, and discuss their properties using some examples. In Section 3, we discuss Thompson sampling and modified Thompson sampling, and extend our setup to allow for targeting based on covariates. Section 4 shows the relative performance of adaptive algorithms in simulations of the proposed sampling method that use data from published RCTs. Section 5 provides a theoretical analysis of the behavior of the proposed modified Thompson algorithm and Section 6 discusses inference for adaptive designs. In Section 7 we describe an RCT where we implemented our approach, demonstrating its practical feasibility [TBD], and section 8 concludes. The supplementary appendix provides additional simulation results and examples of optimal assignments.

# 2 Optimal adaptive design without covariates

Consider a policymaker who would like to maximize the expected value of a binary outcome variable, i.e., a success rate. She has to choose between three or more different policies – or treatments – and she can use an experiment that proceeds in multiple waves. At the end of each experimental wave, outcomes are observed, and treatment assignment in subsequent waves can be based on these observed outcomes. After the experiment concludes, a treatment is chosen for large-scale implementation. We would like to optimally design the experiment for this policy choice problem.

## 2.1 Setup

**Treatments and potential outcomes.** The experiment takes place in waves $t = 1, \dots, T$. Each wave $t$ is a new random draw of $N_t$ experimental units $i = 1, \dots, N_t$ from the population of interest (so the waves are repeated cross-sections).

Each person or unit $i$ in period $t$ can receive one of $k$ different treatments $D_{it} \in \{1, \dots, k\}$, resulting in a binary outcome $Y_{it} \in \{0, 1\}$. Outcome $Y_{it}$ is determined by the potential outcome equation

$$Y_{it} = \sum_{d=1}^{k} \mathbf{1}(D_i = d) \cdot Y_{it}^d.$$

This assumption implies in particular that there is no interference, i.e., outcomes are not affected by the treatments others receive. Random sampling means that the potential outcome vector $(Y_{it}^1, \dots, Y_{it}^k)$ for unit $i$ in period $t$ is an i.i.d. draw from the population of interest. Each treatment $d$ has a stationary unobserved average potential outcome (also known as average structural function)

$$\theta^d = E[Y_{it}^d].$$

**Treatment assignment and state space during the experiment.** Denote by $n_t^d = \sum_i \mathbf{1}(D_{it} = d)$ the number of units assigned to treatment $d$ in wave $t$. The treatment assignment in wave $t$ is summarized by the vector

$$\boldsymbol{n}_t = (n_t^1, \dots, n_t^k) \quad \text{with } \sum_d n_t^d = N_t.$$

The experimenter's problem is to choose $\boldsymbol{n}_t$ at the beginning of wave $t$.

We call $s_t^d = \sum_i \mathbf{1}(D_{it} = d, Y_{it} = 1)$ the number of successes (outcome $Y_{it} = 1$) among those in treatment group $d$ in wave $t$. The outcome of wave $t$ can be summarized by the vector

$$\boldsymbol{s}_t = (s_t^1, \ldots, s_t^k) \quad \text{where } s_t^d \leq n_t^d,$$

collecting the number of successes in each of the treatment groups in wave $t$. These outcomes are observed at the end of wave $t$. Treatment assignment in wave $t+1$ can depend on the outcomes of waves 1 to $t$ and on a randomization device.

Denote the cumulative versions of these terms by

$$\begin{aligned} m_t^d &= \sum_{t' \leq t} n_{t'}^d & r_t^d &= \sum_{t' \leq t} s_{t'}^d \\ \boldsymbol{m}_t &= (m_t^1, \ldots, m_t^k) & \boldsymbol{r}_t &= (r_t^1, \ldots, r_t^k). \end{aligned}$$

Thus, $m_t^d$ is the total number of units assigned to treatment $d$ in waves 1 through $t$, and $r_t^d$ is the total number of successes among these units. With i.i.d. potential outcomes, all relevant information for the experimenter at the beginning of period $t+1$ is summarized by $\boldsymbol{m}_t$ and $\boldsymbol{r}_t$.

**Policy choice and welfare.** After wave $T$, a policy $d^* \in 1, \ldots, k$ will be chosen and implemented with the objective to maximize the expected average of the outcome $Y$ for the whole (remaining) population of interest, net of the cost of treatment. The per-capita expected social welfare of policy $d$ at the end of the experiment is given by

$$SW(d) = E[\theta^d | \boldsymbol{m}_T, \boldsymbol{r}_T] - c^d,$$

that is, the expected success rate of policy $d$, net of the unit cost $c^d$ of implementing $d$. The optimal policy choice after the experiment is given by

$$d^* = \underset{d}{\text{argmax}} \; SW(d).$$

In this formulation, social welfare does not include the outcomes of participants in the experiment. This implies that treatment assignment in each experimental wave is chosen to maximize learning and optimize the choice after wave $T$.

Excluding the welfare of participants from the optimization problem is justified if

the number of experimental units is small relative to the population of interest. A concern for the welfare of participants could easily be added to our objective function, resulting in a hybrid setting between the bandit problem and the setting considered here.

**Bayesian prior and posterior.** Under our assumptions, $Y^d$ has a Bernoulli distribution with unknown parameter $\theta^d$: $Y^d \sim Ber(\theta^d)$. The experiment and treatment assignment are designed to learn as quickly as possible about the individual success rates $\theta^d$ and choose the treatment with the highest $\theta^d$.

We assume that the policymaker holds prior belief

$$\theta^d \sim Beta(\alpha_0^d, \beta_0^d).$$

The $\theta^d$ are mutually independent across $d$. A special case, and the default for applications later in this paper, is the uniform prior $\boldsymbol{\theta} \sim Uniform\left([0,1]^k\right)$, corresponding to $\alpha_0^d = \beta_0^d = 1$ for all $d$.

After outcomes for periods $1, \ldots, t$ are realized, the posterior distribution with a Beta prior is given by

$$\theta^d | \boldsymbol{m}_t, \boldsymbol{r}_t \sim Beta(\alpha_t^d, \beta_t^d) \qquad \begin{aligned} \alpha_t^d &= \alpha_{t-1}^d + s_t^d \\ &= \alpha_0^d + r_t^d \end{aligned} \qquad \begin{aligned} \beta_t^d &= \beta_{t-1}^d + n_t^d - s_t^d \\ &= \beta_0^d + m_t^d - r_t^d. \end{aligned}$$

Moreover, expected social welfare after period $T$, based on the expected success rate for $d$, is

$$SW(d) = \frac{\alpha_0^d + r_T^d}{\alpha_0^d + \beta_0^d + m_T^d} - c^d.$$

From the perspective of the experimenter, the outcomes of wave $t$ after making the treatment assignment decision $\boldsymbol{n}_t$ are subject to two sources of uncertainty: the uncertainty about $\boldsymbol{\theta}$, given $\boldsymbol{m}_{t-1}$ and $\boldsymbol{r}_{t-1}$, and the sampling uncertainty over the distribution of $\boldsymbol{s}_t$, given $\boldsymbol{\theta}$ and $\boldsymbol{n}_t$. The former is given by the $Beta(\alpha_{t-1}^d, \beta_{t-1}^d)$ distribution, the latter by Binomial distributions with parameters $n_t^d$ and $\theta^d$. Integrating out the unknown parameter $\boldsymbol{\theta}$, the number of successes for each treatment in wave $t$ follows a

Beta-Binomial distribution,

$$P(s_t^d = s|\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, n_t^d) = E[P(s_t^d = s|\theta^d, n_t^d)|\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, n_t^d] \\ = \binom{n_t^d}{s} \frac{B(\alpha_{t-1}^d + s, \beta_{t-1}^d + n_t^d - s)}{B(\alpha_{t-1}^d, \beta_{t-1}^d)}. \tag{2.1}$$

## 2.2 Optimal assignment

The choice of treatment assignment $\boldsymbol{n}_t$ for each $t = 1, \ldots, T$ is a dynamic stochastic optimization problem that can in principle be solved using backward induction.[1]

The state at the end of wave $t-1$ is given by $(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1})$, and the action in $t$ is given by $\boldsymbol{n}_t$. The transition between states is described by $\boldsymbol{m}_t = \boldsymbol{m}_{t-1} + \boldsymbol{n}_t$, $\boldsymbol{r}_t = \boldsymbol{r}_{t-1} + \boldsymbol{s}_t$, where the success probabilities are given by Equation (2.1).

Denote by $V_t$ the value function after completion of wave $t$, that is, expected welfare assuming that all future treatment assignment decisions will be optimal, and that the optimal policy is implemented after the experiment. $V_t$ is a function of the state $(\boldsymbol{m}_t, \boldsymbol{r}_t)$. After the experiment is concluded, the value function is given by the optimal choice of policy, based on current beliefs:

$$V_T(\boldsymbol{m}_T, \boldsymbol{r}_T) = \max_d \left(E[\theta^d|\boldsymbol{m}_T, \boldsymbol{s}_T] - c^d\right) = \max_d \left(\frac{\alpha_0^d + r_T^d}{\alpha_0^d + \beta_0^d + m_T^d} - c^d\right). \tag{2.2}$$

Denote by $U_t$ expected welfare at the beginning of wave $t$ when treatment assignment is $\boldsymbol{n}_t$, assuming all future assignment decisions will be optimal:

$$U_t(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t) = \sum_{\boldsymbol{s}:\boldsymbol{s}\leq\boldsymbol{n}_t} P(\boldsymbol{s}_t = \boldsymbol{s} \,|\, \boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t) V_t\left(\boldsymbol{m}_{t-1} + \boldsymbol{n}_t, \boldsymbol{r}_{t-1} + \boldsymbol{s}\right), \tag{2.3}$$

where the probabilities for each vector of successes $P(\boldsymbol{s}_t = s|\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t)$ is determined by the Beta-Binomial distribution of Equation (2.1). The period $t$ value function and the optimal experimental design satisfy

$$V_{t-1}(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}) = \max_{\boldsymbol{n}_t:\, \sum_d n_t^d \leq N_t} U_t(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t) \\ \boldsymbol{n}_t^*(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}) = \operatorname*{argmax}_{\boldsymbol{n}_t:\, \sum_d n_t^d \leq N_t} U_t(\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{n}_t). \tag{2.4}$$

[1] In practice, computational challenges motivate non-optimal but tractable alternatives, as we will discuss below.

Together, these equations define a solution for the experimental design problem.

## 2.3 Optimal design in a simple example

In this section, we discuss optimal experimental designs in a simple example, in order to build intuition and to provide motivation for our proposed modified Thompson sampling procedure below. Suppose we have ten experimental units that we can enroll in two waves. There are three treatments. The cost of all treatments is the same, so we set $\boldsymbol{c} = 0$ for simplicity. We impose a uniform prior for $\boldsymbol{\theta}$.

**Dividing the sample between first and second wave.** A first question to consider is how to divide the total sample of 10 units between the two waves. For each division $(N_1, 10 - N_1)$ between the two waves, we can calculate expected welfare $V_0$ at the outset of wave 1, using the value function derived in Section 2.2.

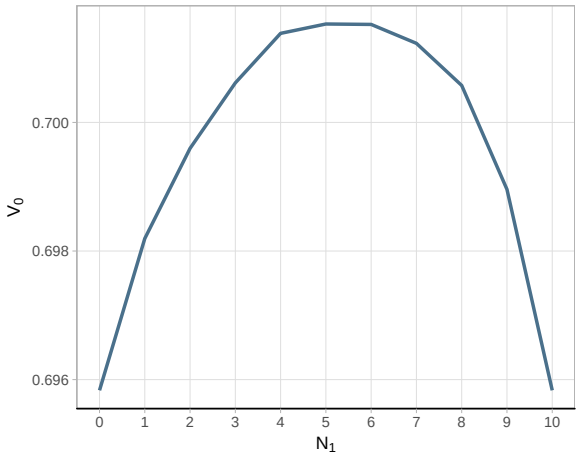
Figure 2.1 plots expected welfare as a function of the sample size $N_1$ in wave 1. The boundary cases $N_1 = 0$ and $N_1 = 10$ correspond to an experiment with only one wave. The optimal split assigns either 5 or 6 units to the first wave. Splitting the sample in this manner allows to learn from the first-wave assignment (e.g. of two units per treatment if $N_1 = 6$) and then focus attention on the treatments with higher values in the second wave.[2]

**Assigning treatments.** Based on Figure 2.1, we set $N_1 = 6$, so that we have $N_2 = 4$ in the second wave. Driven by the symmetric prior, it is optimal to assign 2 units to each of the 3 treatments in wave 1. Optimal assignment in wave 2 depends on the outcomes of the first wave. We explore several scenarios in Figure 2.2.[3] This figure plots expected welfare for any second-wave treatment assignment in the simplex $n_2^1 + n_2^2 + n_2^3 = 4$, conditional on first-wave outcomes. For each scenario, the number of successes in each treatment in the first wave determines the prior for treatment assignments in the second wave. Our uniform prior for $\boldsymbol{\theta}$ implies a Beta posterior, where for $s_1^d \in \{0, 1, 2\}$ we get $\alpha_1^d = 1 + s_1^d$ and $\beta_1^d = 1 + 2 - s_1^d$. This Beta posterior has a mean of $(1 + s_1^d)/4$. Figures A5 and A6 in the Appendix complement our analysis of these scenarios by plotting the mapping from first wave outcomes to optimal second wave treatment assignments.

[2]The welfare differences across alternative designs are relatively small in this setting, owing to the small number of units involved. In our simulations calibrated to more realistic settings below we will find quantitatively important differences.

[3]The figures in Appendix A.3 explore similar scenarios for different sizes of wave $N_2$.

Figure 2.1: Dividing the sample across waves.

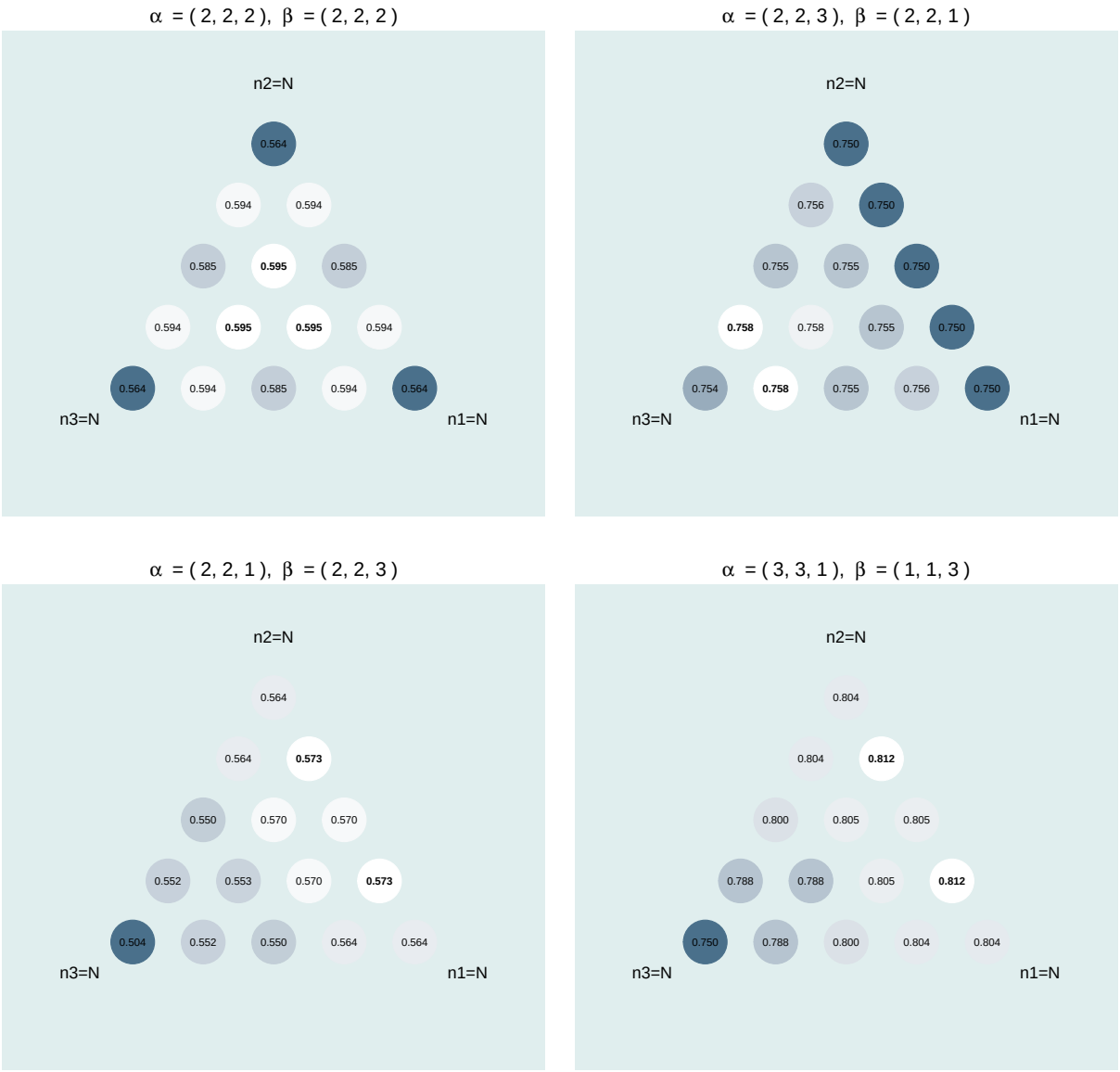



**Notes:** The graph shows expected welfare $V_0$ as a function of the sample size $N_1$ in period 1, assuming a total sample size of 10 and three treatments, for a uniform prior.

**Four scenarios** The four scenarios we consider are $\boldsymbol{s}_1 = (1,1,1)$, $\boldsymbol{s}_1 = (1,1,2)$, $\boldsymbol{s}_1 = (1,1,0)$, and $\boldsymbol{s}_1 = (2,2,0)$. In the first scenario, each treatment had one success and one failure, leading to a posterior that is again symmetric across treatments. In this scenario, shown in the top left of Figure 2.2, it is optimal to assign 2 units to either of the three treatments, and 1 unit to the other two arms. In the second scenario, treatment 3 performed better than treatments 1 and 2. In this scenario, shown in the top right of Figure 2.2, it is optimal to assign 3 units to treatment 3, and 1 unit to either of the other two arms. In the third and fourth scenario, treatment 3 performed worse than treatments 1 and 2. In these scenarios, shown in the bottom part of Figure 2.2, it is optimal to assign no units to treatment 3, 3 units to either of treatment 1 or 2, and 1 to the other. Interestingly, this dominates (though not by much) the assignment of 2 units to each of treatment 1 and 2.

**Discussion** These examples show that dividing the sample equally between treatment arms is generally not optimal. Moreover, in each example, the largest number of units is assigned to the treatment arms with the highest expected return. This reflects that more precise effect estimates for treatment arms with low expected return are unlikely to affect the ultimate policy decision. This is true even though our objective function does not assign any weight to the welfare of experimental units; there is no exploitation motive. This feature of optimal treatment assignments is approximated by Thompson sampling and modified Thompson sampling, which will be introduced in

Figure 2.2: Expected welfare as a function of treatment assignment




**Notes:** This figure shows the expected welfare $U_2$ for each possible treatment assignment $\boldsymbol{n}_2 = (n_2^1 + n_2^2 + n_2^3)$ in wave 2 (which is of size 4), taking as given the Beta-prior parameters $\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1$ which were determined by the outcomes of wave 1 (which is of size 6). For example, the upper right panel is for the case where treatment 1 and 2 each had one success, but treatment 3 had 2 successes. Note that the color scaling differs across the plots for better readability.

Table 1: Two decision problems

| | **Estimation** | **Policy choice** |
|---|---|---|
| Decision | Estimate $\hat{\boldsymbol{\theta}}$ | Policy $d^* = \text{argmax}_d \, E[\theta^d \vert \boldsymbol{m}_T, \boldsymbol{r}_T]$ |
| Loss function | $\sum_d (\hat{\theta}^d - \theta^d)^2$ | $\max_d \theta^d - \theta^{d^*}$ |
| Risk function | $\sum_d \text{Var}_\theta(\hat{\theta}^d) + \text{Bias}^2_\theta(\hat{\theta}^d)$ | $\max_d \theta^d - E_\theta[\theta^{d^*}]$ |
| Bayes risk | $\sum_d E[(\hat{\theta}^d - \theta^d)^2]$ | $\max_d \theta^d - E[\max_d E[\theta^d \vert \boldsymbol{m}_T, \boldsymbol{r}_T]]$ |

Section 3 below.

A last observation is that even with symmetric priors, a symmetric assignment is not necessarily optimal. Consider for instance the case $\alpha = (3, 3, 1)$, $\beta = (1, 1, 3)$. The prior distribution for treatments 1 and 2 is the same. The optimal design, however, assigns either more units to treatment 1 or to 2. This reflects a non-convexity in the value of information, due to the concave objective function $\max_d \big(E[\theta^d \vert \boldsymbol{m}_T, \boldsymbol{s}_T] - c^d\big)$. This situation is analogous to option pricing, where higher volatility can increase the value of a stock option which is only exercised for high profit realizations.

## 2.4 Comparison with alternative design problems

The goal of the experiments considered in this paper is to inform a policy choice after the experiment. This contrasts with two alternative goals, (i) estimating treatment effects, and (ii) maximizing the outcomes of experimental units. It is useful to compare the objective functions of these alternative settings to our objective.

**Estimation.** In our experiment, the goal is to choose a policy in order to maximize expected outcomes. A more common goal of experimental design is to obtain estimates that are as precise as possible, e.g. of average treatment effects. Table 1 provides a comparison of the two goals, where we assume for simplicity that $\boldsymbol{c} = 0$.

In the case of estimation, the goal is to obtain a vector of estimates $\hat{\boldsymbol{\theta}}$ after the experiment, e.g. by calculating simple sample averages, or a Bayesian posterior mean. A common way to evaluate estimators is based on quadratic error loss. If we take the *ex ante* expectation of loss given the true parameter $\boldsymbol{\theta}$, we obtain the risk function, and for unbiased estimators, this is simply the estimator variance. Averaging over the prior distribution for $\boldsymbol{\theta}$ gives Bayes risk.

Analogously, in the case of policy choice, the goal is to choose a policy $d^*$ for large-scale implementation. The loss function is the difference between the optimal

treatment – sometimes called the oracle-optimal choice – and the chosen treatment. It is also called the regret function and defined as Regret $= \max_d \theta^d - \theta^{d^*}$. The risk function, or expected regret, is given by $\max_d \theta^d - E_\theta[\theta^{d^*}]$.

It is useful to keep the analogy between estimator variance and expected regret in mind for our subsequent discussions. The difference between these two objectives (minimizing estimator variance versus minimizing expected regret) drives the difference in the resulting design recommendations. The analogy between them is useful when considering practical questions such as the choice of sample size.

**Bandit problems versus policy choice.** Another experimental design problem that has received much attention is the multi-armed bandit problem. In the basic multi-armed bandit problem, experimental units arrive one by one. Each unit is assigned a treatment, and the outcome is observed before the next unit arrives. The objective is to maximize the (discounted) average outcome of experimental units.

The setup we consider differs in two ways from such bandit problems. First, and somewhat less importantly, we consider units arriving in waves ("batched" assignment, in machine learning terminology), and assume that the experiment stops after a pre-determined number of waves.

Second, and crucially, in the bandit setting the experimenter cares about the (discounted) outcomes or welfare of the treated units themselves. This results in a tradeoff between exploitation (assigning the current best treatment) and exploration (learning about other treatments in order to make better choices for units arriving in the future). A common intuition for bandit problems suggests that the exploitation motive favors treatments with the best past outcomes, while the exploration motive favors balanced assignment. In our setting, there is no exploitation motive. Nonetheless, there is a strong rationale to focus on the highest-performing treatments, and it hinges on the objective of learning (only) about the welfare-maximizing policy choice. By contrast, if the objective is to learn about all treatments, such as in the estimation problem, this is not a property of the optimal design.

# 3 Modified Thompson sampling

**Computational cost of optimal solutions** Based on the Equations (2.4), we can in principle solve for the optimal treatment assignment using dynamic programming. This involves brute-force enumeration of all possible outcomes and actions, where both have finite support. We did this above for some illustrative examples. With larger sample sizes and a greater number of treatments, however, solving for the optimal assignment quickly becomes infeasible. To see this, consider the problem of finding $\boldsymbol{n}_t^*$, assuming knowledge of $V_t$. For each $\boldsymbol{n}_t$, evaluating $U_t$ requires enumerating all possible realizations of $\boldsymbol{s}_t$, of which there are $\prod_d n_t^d = O(N_t^k)$. For each realization of $\boldsymbol{s}_t$, we need to evaluate the Beta-Binomial likelihood. Then we need to calculate $U_t$ for each possible $\boldsymbol{n}_t$; there are $\binom{N_t+k-1}{k-1} = O(N_t^k)$ such possible assignment vectors. The required computation time is thus of order $O(N_t^{2k})$, times the computation time for $V_t$ and for the Beta-Binomal likelihood.

An alternative to full optimization is the use of simpler algorithms that approximate the optimal solution. Such approximate algorithms are widely used for bandit problems, such as online dynamic experiments, for instance in the placement of ads. One of the most popular (and oldest) such algorithms is so-called Thompson sampling, originally proposed by Thompson (1933) in the context of clinical trials.

## 3.1 Thompson sampling and modified Thompson sampling

**Thompson sampling** Consider a special case of the setting described in Section 2.1 above where each wave is of size 1, so that units arrive sequentially (and we can drop the subscript $i$). In each period $t$, assign any treatment $d$ with probability equal to the posterior probability, given past outcomes, that it is in fact the optimal treatment,

$$p_t^d = P(D_t = d|\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}) = P(d = \underset{d'}{\text{argmax}}\ (\theta^{d'} - c^{d'})|\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}).$$

This prescription is easy to implement, by sampling just one draw $\hat{\theta}_t$ from the posterior given $\boldsymbol{m}_{t-1}$ and $\boldsymbol{r}_{t-1}$, and setting

$$D_t = \underset{d}{\text{argmax}}\ (\hat{\theta}_t^d - c^d).$$

In the context of the Beta-Binomial model outlined above, $\hat{\theta}_t$ is sampled from its Beta posterior. Thompson sampling can also be applied in much more general settings, with more complicated policy spaces, prior distributions, and data likelihoods. An excellent overview can be found in Russo et al. (2018). This approach is easily adapted to batched settings such as ours, where posteriors are based on the outcomes of all preceding waves.

**Modified Thompson sampling** We can improve on standard Thompson sampling in our context. We propose the following two modifications. These modifications improve performance in our simulations. We also prove analytically that they improve expected welfare in large samples. The **first modification** that we propose is designed to reduce randomness in the treatment assignment. Rather than drawing each $D_{it}$ independently from the distribution $(p_t^1, \ldots, p_t^k)$, we assign a non-random share $p_t^d$ (up to required rounding) of observations in wave $t$ to treatment $d$. We will refer to treatment assignment based on this modification alone as *expected Thompson sampling*.

The **second modification** that we propose replaces the assignment probabilities $(p_t^1, \ldots, p_t^k)$ with the following transformed probabilities:

$$q_t^d = S_t \cdot p_t^d \cdot (1 - p_t^d)$$
$$S_t = \frac{1}{\sum_d p_t^d \cdot (1 - p_t^d)}.$$

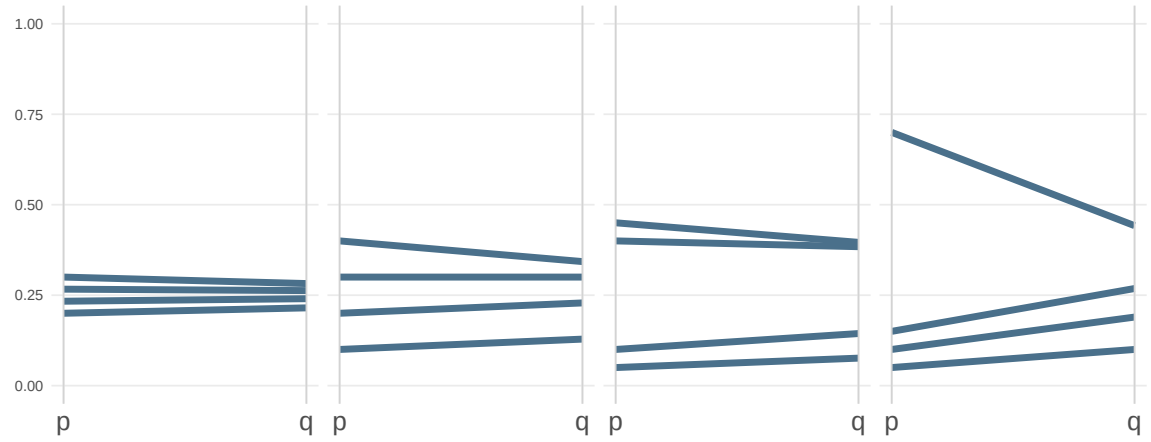
We will analyze this modification and its justification in detail in Section 5 below. We will refer to treatment assignment based on both of these modifications as *modified Thompson sampling*. To get some initial intuition for the mapping from the vector of probabilities $\boldsymbol{p}_t$ to the transformed vector $\boldsymbol{q}_t$, Figure 3.1 plots a few examples.

Relative to standard Thompson sampling, this modification improves expected welfare by shifting observational units from the best-performing treatment to its close competitors, thereby improving power for rejecting the latter when choosing the policy $d^*$ after conclusion of the experiment.

## 3.2 Covariates, targeted assignment, and hierarchical priors

The model introduced in Section 2.1 did not involve any covariates. Suppose now that we additionally observe a covariate $X_i$ with finite support for each unit $i$, before assigning a treatment $D_i$. Assume furthermore that the optimal treatment assignment

Figure 3.1: Illustration of modified Thompson probabilities




**Notes:** This figure shows examples of the mapping from the vector of Thompson probabilities $\boldsymbol{p}_t$ to the vector of modified Thompson probabilities $\boldsymbol{q}_t$.

policy chosen by the policymaker at the end of the experiment might also condition treatment on $X$, assigning $d^*(x)$ to units characterized by $X = x$. We discuss three approaches to treatment assignment in this context, (i) considering each stratum as a separate experiment, (ii) (modified) Thompson sampling based on a hierarchical prior, and (iii) (modified) Thompson sampling based on an empirical Bayes approach.

**Considering each stratum as separate experiment** How does this affect our analysis? One possibility would be to simply treat each stratum, defined by a value of $X$, as a separate experiment. Within strata, the previous analysis then applies almost verbatim. Note, however, that we might not know in advance the covariate values for future waves. This implies that the strata-specific sample sizes are random ex-ante, thus adding an additional dimension of uncertainty to our optimization problem. Additionally, and more interestingly, we might want to leverage information from some strata in order to learn something about average potential outcomes for other strata. For example, if a medical treatment works well for patients aged 50-60 years, that might suggest that it also works well for patients aged 60-70 years.

**Hierarchical priors** Formally, this consideraton translates into a prior with statistical dependence between the vectors of average potential outcomes $\boldsymbol{\theta}^x = (\theta^{1x}, \ldots, \theta^{kx})$ across different strata $x$. One natural way to model such dependence is by constructing a hierarchical model, cf. Chapter 5 in Gelman et al. (2014). The following describes

the approach we use in our simulations and applications using covariates below.

Consider the following generalization of the setting considered thus far, where $x$ indexes strata, and $d$ denotes treatment values. Let $\theta^{dx}$ be the corresponding average potential outcome. Assume

$$\begin{aligned} Y_{it}^d | X_{it} = x, \theta^{dx}, (\alpha_0^d, \beta_0^d) &\sim Ber(\theta^{dx}) \\ \theta^{dx} | (\alpha_0^d, \beta_0^d) &\sim Beta(\alpha_0^d, \beta_0^d) \\ (\alpha_0^d, \beta_0^d) &\sim \pi, \end{aligned}$$

where $\pi$ is some prior distribution for $(\alpha^d, \beta^d)$, and parameters are independent across treatment arms. It follows immediately that $s_t^{dx}$, the number of successes for treatment $d$ and stratum $x$, follows a Beta-Binomial distribution given $(\alpha^d, \beta^d)$ and $n_t^{dx}$.

Intuitively, updating based on this prior works as follows. For each treatment $d$, consider the success rates $s_t^{dx}$ across different strata $x$. Based on these success rates, learn the mean and dispersion of $\theta^{dx}$ across strata, as reflected in $(\alpha^d, \beta^d)$. Then use these as a prior, in conjunction with $s_t^{dx}$ for a given stratum $x$, to learn about $\theta^{dx}$ for that stratum.

**Markov Chain Monte Carlo** More formally, and in a slight change of notation relative to our baseline model, denote by $\boldsymbol{\theta}, \boldsymbol{m}_t, \boldsymbol{r}_t$ the vectors of parameters, cumulative trials and successes indexed by both $d$ and $x$. Let $\rho$ index replication draws, with $\rho$ ranging from 1 to $R$. We sample from the posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ given $\boldsymbol{m}_{t-1}, \boldsymbol{r}_{t-1}$ using the following Markov Chain Monte Carlo algorithm.

1. Gibbs step:
   Given $\boldsymbol{\alpha}_{\rho-1}$ and $\boldsymbol{\beta}_{\rho-1}$, draw $\theta^{dx}$ from the $Beta(\alpha_{\rho-1}^d + s^{dx}, \beta_{\rho-1}^d + m^{dx} - s^{dx})$ distribution.

2. Metropolis steps:

   - Given $\boldsymbol{\beta}_{\rho-1}$ and $\boldsymbol{\theta}_\rho$, draw $\alpha_\rho^d$ by sampling from a normal proposal distribution (truncated below), and accept this draw if an independent uniform draw is less than the ratio of the posterior for the new draw, relative to the posterior for $\alpha_{\rho-1}^d$. Otherwise set $\alpha_\rho^d = \alpha_{\rho-1}^d$.
   - Similarly for $\boldsymbol{\beta}_{\rho-1}$ given $\boldsymbol{\theta}_\rho$ and $\boldsymbol{\alpha}_\rho$.

Markov Chain Monte Carlo methods are reviewed in Gelman et al. (2014), chapter 11. This algorithm converges to a stationary distribution that equals the joint posterior of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ given $\boldsymbol{m}_t, \boldsymbol{r}_t$. In particular, we have that the posterior probability of a treatment $d$ being optimal given $x$ is given by

$$P\left(d = \underset{d'}{\text{argmax}}\ \theta^{d'x} | \boldsymbol{m}_t, \boldsymbol{r}_t\right) = \underset{R\to\infty}{\text{plim}}\ \frac{1}{R}\sum_{\rho=1}^{R} \mathbf{1}\left(d = \underset{d'}{\text{argmax}}\ \theta_{\rho}^{d'x}\right).$$

**Empirical Bayes** A third possibility, which is closely related to the hierarchical Bayes approach, is the use of an empirical Bayes approach. As before, assume that

$$\begin{aligned} Y_{it}^d | X_{it} = x, \theta^{dx}, (\alpha_0^d, \beta_0^d) &\sim Ber(\theta^{dx}) \\ \theta^{dx} | (\alpha_0^d, \beta_0^d) &\sim Beta(\alpha_0^d, \beta_0^d). \end{aligned}$$

For the empirical Bayes approach, we do not require a prior distribution over $(\alpha_0^d, \beta_0^d)$. Instead, we generate "posterior" draws as follows:

1. First, for each $d$ estimate $(\alpha^d, \beta^d)$ using **maximum likelihood**, based on the Beta-Binomial likelihood for the observed $\boldsymbol{r}_{t-1}^d$.

2. For each draw of $(\alpha^d, \beta^d)$, sample a draw of $\boldsymbol{\theta}^d$ from its Beta posterior given $(\alpha^d, \beta^d)$ and given $\boldsymbol{m}_{t-1}^d, \boldsymbol{r}_{t-1}^d$.

Then proceed as before. The empirical Bayes approach works particularly well when there are many strata, so that $\alpha_0^d$ and $\beta_0^d$ can be learned with high confidence.

# 4 Calibrated simulations

We next present simulation evidence on the performance of alternative treatment assignment algorithms, using parameter vectors $\boldsymbol{\theta}$ and sample sizes $M_T$ calibrated to data from published experiments in development economics. The purpose of calibration is to "tie our hands" in choosing designs for our simulations. We thus opted for simplicity, rather than realism, in the assumptions driving our calibrations.

**Experiments from the literature** We consider the experiments discussed in Ashraf et al. (2010), Bryan et al. (2014), and Cohen et al. (2015).

Ashraf et al. (2010) conducted a field experiment in Zambia involving Clorin, a disinfectant. During a door-to-door sale of Clorin to about 1,000 households in Lusaka, each participating household was offered a bottle of Clorin for a randomly chosen offer price, at or below the retail price. The treatment in this experiment is the price offered, ranging from 300 to 800 Zambian Kwacha. The outcome is whether the household bought the bottle of Clorin.

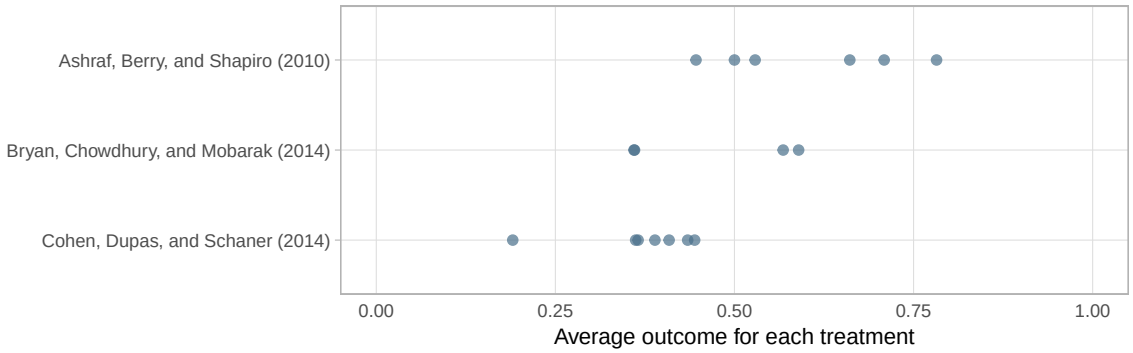
Bryan et al. (2014) conducted a field experiment in rural Bangladesh. Households were randomly assigned a cash or credit incentive of $8.50, conditional on a household member migrating during the 2008 monga (lean) season. This amount covers the round-trip travel cost. The treatments in this experiment are cash, credit, information, and a control group. The outcome is whether at least one household member migrated.

Cohen et al. (2015) conducted a field experiment in three districts of Western Kenya. Households were randomly assigned one of three subsidy levels for the purchase of artemisinin combination therapies (ACT), an antimalarial drug. They were also randomly offered a rapid detection test (RDT) for malaria. The treatments in this experiment are 3 subsidy levels with or without RDT, and a control group. The outcome is whether the household actually bought ACT.

## 4.1 No covariates

We first consider simulations without covariates, corresponding to the setting discussed in Section 2.1. Throughout, we make two simplifying assumptions, to tie our hands and keep our analysis transparent. First, we ignore clustering in the sampling and treatment assignment of the original experiments. Second, we assume that the policymaker's goal is to maximize the average of the measured outcome, and that the cost of each

Figure 4.1: Average outcomes across treatment arms in published experiments




treatment is the same, w.l.o.g. $\boldsymbol{c} = 0$. This is of course unrealistic, but allows us to avoid ad-hoc assumptions regarding $\boldsymbol{c}$, and to focus on the benefit of adaptive assignment for a range of parameter vectors $\boldsymbol{\theta}$.

**Calibrated parameter values** Figure 4.1 shows the average outcomes across treatment arms for each of the three experiments. We set the vectors $\boldsymbol{\theta}$ equal to these average outcomes, for the purpose of our simulations. These vectors show interesting differences across the three experiments, which will be relevant for understanding the results of our simulations.

For Ashraf et al. (2010), there are roughly evenly spaced average outcomes ranging from .44 to .78 across 6 treatments. This is a setting where it is comparatively easy to statistically detect which treatments are performing better, so that we would expect benefits of adaptation even for moderate sample sizes.

For Bryan et al. (2014), there are two worse treatments with average outcomes of about .36 (these two treatments are indistinguishable in Figure 4.1), and two better treatments that are very close, with average outcomes of .57 and .59. In this setting, it is easy to detect which two treatments perform better. Among these two, however, it takes a large amount of information to figure out which is the best. The returns of finding the best treatment among the top two, on the other hand, are not very large.

For Cohen et al. (2015), the top 6 treatments are again roughly evenly spaced, with average outcomes for these ranging from .36 to .44 (the second and third treatment from the bottom are again indistinguishable in Figure 4.1). This setting is similar to Ashraf et al. (2010), except that the best treatments are closer and thus harder to distinguish.

**Algorithms** We compare four different algorithms. The first algorithm, which serves as a benchmark, is *non-adaptive* and assigns an equal share of units to each of the treatment arms. This is the conventional recommendation for experimental design. The second algorithm is standard *Thompson* sampling. The third algorithm, *expected Thompson*, assigns a non-random share of units in each wave based on the Thompson probabilities. The fourth algorithm, our preferred approach, is *modified Thompson* sampling as described in Section 3.

**Performance criteria** We evaluate the performance of these algorithms in terms of the distribution of regret across 20,000 simulation draws. Since we set $\boldsymbol{c} = 0$ (or constant across treatment arms), regret is given by the difference between the welfare generated by the optimal treatment, and welfare for the policy $d^*$ with the highest posterior mean after conclusion of the experiment. That is,

$$d^* = \underset{d}{\text{argmax}}\ E[\theta^d | \boldsymbol{m}_T, \boldsymbol{r}_T],$$
$$\text{regret} = \max_d \theta^d - \theta^{d^*}.$$

For each of our simulations the vector $\boldsymbol{\theta}$ is fixed across simulation draws, and thus the same holds for $\max_d \theta^d$, so that (in this context) average regret is just a convenient renormalization of the average of welfare $\theta^{d^*}$. We also report the share among our 20,000 simulation draws for which the optimal treatment was chosen after conclusion of the experiment, that is for which regret $= 0$.

**Simulation results** The tables and figures on the following pages show our simulation results for these settings. These results are based on total sample sizes equal to the original experiments. Appendix A.2 provides similar results for total sample sizes equal to half the original, and equal to 1.5 times the original sample size.

Both tables and figures describe the distribution of regret across simulations. The tables show the average of regret, and the probability that the optimal policy is chosen (corresponding to regret $= 0$) for each of the four algorithms considered. The figures compare the full distribution of regret between *non-adaptive* assignment and *modified Thompson*. They show the probability mass functions (histograms) and the quantile functions of the distribution of regret. Note in particular that a uniformly lower quantile function for modified Thompson sampling, relative to non-adaptive assignment,

implies that its distribution of regret is first-order stochastically dominated. The integrated difference between the two quantile functions equals the decrease in average regret (increase in average welfare) that we gain from switching to modified Thompson sampling.

There are several noticeable patterns across the simulations.

1. Modified Thompson sampling, as proposed in this paper, consistently performs better than expected Thompson sampling, which performs very similarly to Thompson sampling, and all of these outperform non-adaptive assignment.

2. Adaptive designs with more waves consistently outperform designs with fewer waves (for the same total sample size).

3. The gains from adaptive design in terms of average regret are largest in the application to Ashraf et al. (2010), followed by Cohen et al. (2015). The gains for Bryan et al. (2014) are somewhat smaller.

   The gains in the probability of choosing the optimal treatment are even more pronounced.

The figures reveal the following additional properties of modified Thompson sampling.

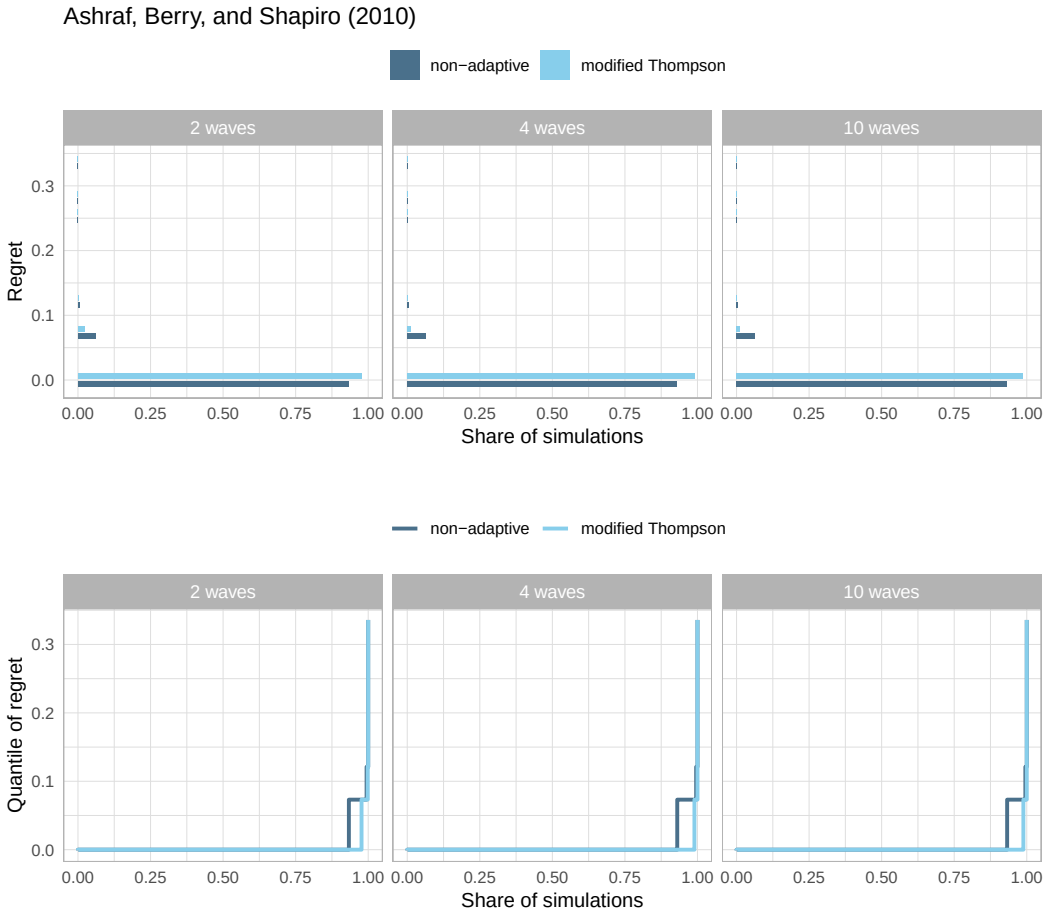
4. The probability of choosing the best treatment is strictly larger than under non-adaptive assignment, for every setting considered.

5. More generally, the distribution of regret under modified Thompson sampling first-order stochastically dominates the corresponding distribution under non-adaptive assignment.

6. For Ashraf et al. (2010) and Bryan et al. (2014), both approaches pick one of the best two treatments with high probability. For Cohen et al. (2015), the distribution is more dispersed, owing to smaller treatment differences.

Table 2: Ashraf, Berry, and Shapiro (2010)

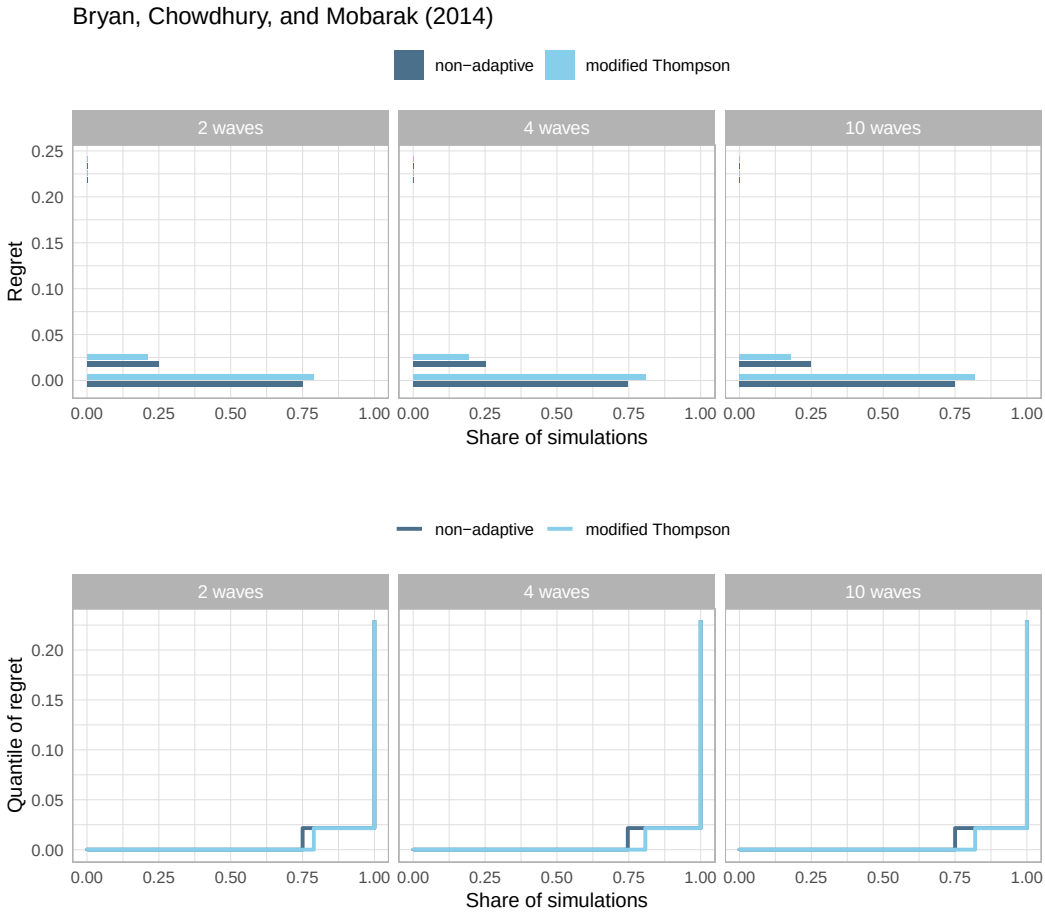
| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.002 | 0.001 | 0.001 |
| expected Thompson | 0.002 | 0.001 | 0.001 |
| Thompson | 0.002 | 0.001 | 0.001 |
| non-adaptive | 0.005 | 0.005 | 0.005 |
| Share optimal | | | |
| modified Thompson | 0.977 | 0.990 | 0.988 |
| expected Thompson | 0.970 | 0.981 | 0.983 |
| Thompson | 0.971 | 0.981 | 0.983 |
| non-adaptive | 0.933 | 0.930 | 0.932 |
| Units per wave | 502 | 251 | 100 |




**Notes:** The table shows average regret and the share of replications for which the optimal treatment was chosen across 20,000 simulation replications. The histograms and quantiles show the distribution of regret. Parameters are calibrated based on the data of Ashraf et al. (2010). Total sample size is equal to the original sample size.

Table 3: Bryan, Chowdhury, and Mobarak (2014)

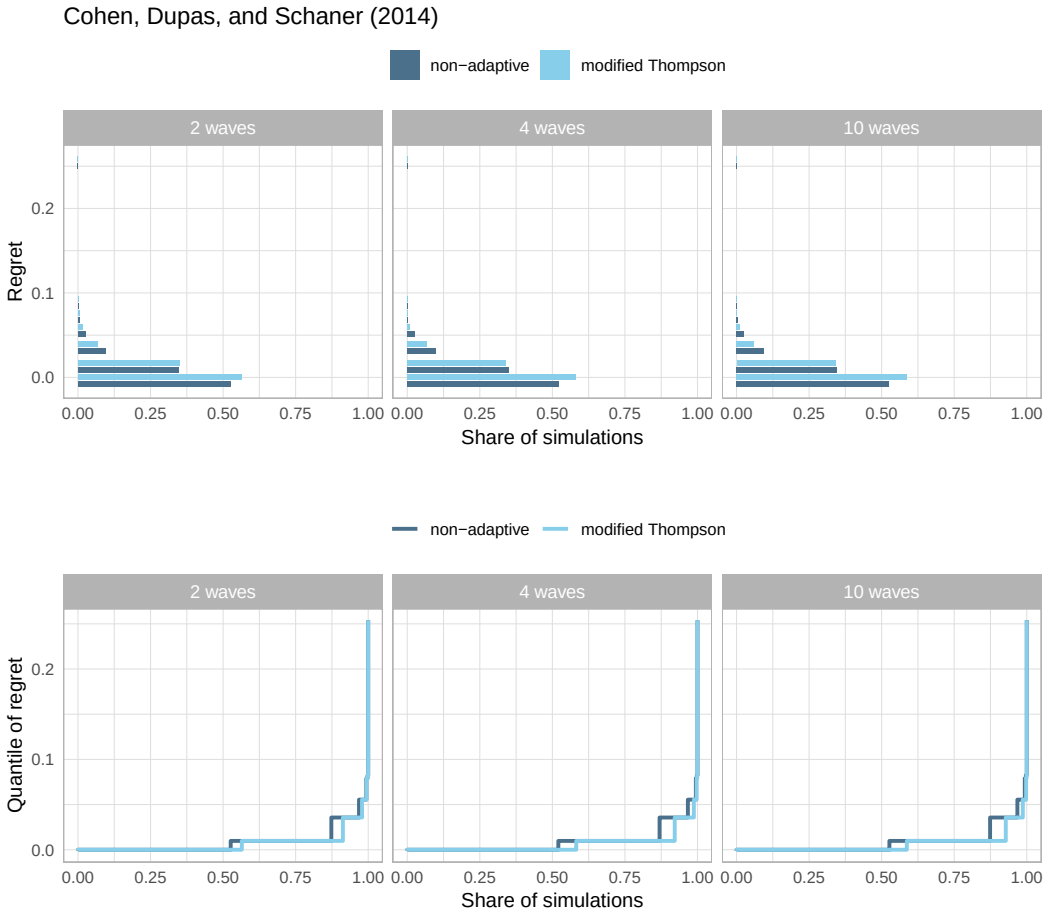
| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.005 | 0.004 | 0.004 |
| expected Thompson | 0.005 | 0.004 | 0.004 |
| Thompson | 0.005 | 0.004 | 0.004 |
| non-adaptive | 0.005 | 0.005 | 0.005 |
| Share optimal | | | |
| modified Thompson | 0.789 | 0.807 | 0.820 |
| expected Thompson | 0.784 | 0.800 | 0.804 |
| Thompson | 0.786 | 0.796 | 0.808 |
| non-adaptive | 0.750 | 0.747 | 0.750 |
| Units per wave | 935 | 467 | 187 |




**Notes:** The table shows average regret and the share of replications for which the optimal treatment was chosen across 20,000 simulation replications. The histograms and quantiles show the distribution of regret. Parameters are calibrated based on the data of Bryan et al. (2014). Total sample size is equal to the original sample size.

Table 4: Cohen, Dupas, and Schaner (2014)

| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.007 | 0.006 | 0.006 |
| expected Thompson | 0.007 | 0.006 | 0.006 |
| Thompson | 0.007 | 0.007 | 0.006 |
| non-adaptive | 0.009 | 0.009 | 0.009 |
| Share optimal | | | |
| modified Thompson | 0.565 | 0.582 | 0.587 |
| expected Thompson | 0.564 | 0.582 | 0.575 |
| Thompson | 0.562 | 0.581 | 0.590 |
| non-adaptive | 0.526 | 0.521 | 0.527 |
| Units per wave | 1080 | 540 | 216 |




**Notes:** The table shows average regret and the share of replications for which the optimal treatment was chosen across 20,000 simulation replications. The histograms and quantiles show the distribution of regret. Parameters are calibrated based on the data of Cohen et al. (2015). Total sample size is equal to the original sample size.

## 4.2 Targeting based on covariates

We next turn to simulations using covariates. Covariates are used both for experimental treatment assignment, and for targeting of $d^*$, corresponding to the setting discussed in Section 3.2. We calibrate both the distribution $p_x$ of units across strata defined by covariate values $x$, and the conditional average potential outcomes $\theta^{dx}$, to the data.

**Calibrated parameter values** We choose the following covariates in order to define strata for targeting. Our choice of these covariates was guided by the requirement of observing units for each combination of covariate and treatment in the original experiments, so that we can calibrate $\theta^{dx}$ to observed averages. For Ashraf et al. (2010), we set $X$ equal to a location indicator, corresponding to one of five low-income peri-urban areas in Lusaka. For Bryan et al. (2014), we set $X$ equal to an indicator for literacy. For Cohen et al. (2015), we set $X$ equal to an indicator for quartiles of distance to clinic.

Figure A1 in the Appendix shows the average outcomes for each treatment and covariate value across the applications considered. We again set the vectors $\boldsymbol{\theta}$ equal to these average outcomes, for the purpose of our simulations. This figure also shows the number of observations in each of the strata, which we use to calibrate the probabilities $p_x$ of each covariate value in the population.

**Algorithms** We compare five different algorithms. The first algorithm is *non-adaptive* assignment *neglecting covariates*. The second algorithm is *non-adaptive* assignment, *stratified* on covariates. In each wave and each stratum, as defined by a value of the covariates, we assign an equal share of units to each of the treatment arms. The third algorithm is *Thompson* sampling, using the hierarchical Bayes posterior discussed in Section 3.2. We use a default prior for the hyper-parameters $(\alpha_0^d, \beta_0^d)$, with density equal to $(\alpha_0^d + \beta_0^d)^{-2.5}$, up to a multiplicative constant. In doing so, we follow the recommendation of Gelman et al. (2014), p110, for picking a "non-informative" hyper-prior. The fourth algorithm is *expected Thompson*. This algorithm uses the posterior probabilities of being optimal, from the hierarchical Bayes posterior, to assign non-random shares of each stratum in each wave to each treatment. The fifth algorithm is *modified Thompson*, which proceeds like expected Thompson but modifies the assignment shares as discussed in Section 3, separately within each stratum.

**Performance criteria** The policies considered now target treatment assignment based on covariates, so that a unit with covariate value $x$ is assigned to $d^*(x)$. Correspondingly, regret in this setting is defined via

$$d^*(x) = \underset{d}{\text{argmax}}\ E[\theta^{dx}|\boldsymbol{m}_T, \boldsymbol{r}_T] - c^d,$$
$$\text{regret} = \sum_x p_x \left( \max_d \theta^{dx} - \theta^{d^*(x)} \right),$$

where $p_x$ is the probability of covariate value $x$ in the population, and $\boldsymbol{m}_T, \boldsymbol{r}_T$ are the cumulative trials and successes for all covariate and treatment combinations. We again report average regret across simulations, as well as the share of simulations for which the optimal policy function $d^*(\cdot)$ was chosen, that is for which regret $= 0$.

**Simulation results** The following tables show our simulation results for these settings, using 10,000 simulation draws. These tables are based on total sample sizes equal to the original experiments. The tables in Appendix A.2 again provide similar results for total sample sizes equal to half the original, and equal to 1.5 times the original sample size.

The following patterns emerge.

1. Again, modified Thompson sampling and Thompson sampling dominate non-adaptive assignment, whether stratified or fully random.

2. Regret is larger than in the corresponding settings without covariates. This is due to the fact that we are faced with the harder problem of figuring out the optimal treatment for each stratum, rather than just the unconditionally optimal treatment.

3. Correspondingly, the share of simulations for which the optimal policy was chosen is smaller.

4. More waves again result in lower average regret for the adaptive designs.

Table 5: Ashraf, Berry, and Shapiro (2010)

| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.013 | 0.011 | 0.011 |
| expected Thompson | 0.013 | 0.011 | 0.011 |
| Thompson | 0.013 | 0.012 | 0.011 |
| non-adaptive stratified | 0.017 | 0.017 | 0.017 |
| non-adaptive | 0.024 | 0.019 | 0.017 |
| Share optimal | | | |
| modified Thompson | 0.172 | 0.196 | 0.205 |
| expected Thompson | 0.176 | 0.201 | 0.205 |
| Thompson | 0.176 | 0.195 | 0.212 |
| non-adaptive stratified | 0.134 | 0.141 | 0.135 |
| non-adaptive | 0.096 | 0.127 | 0.141 |
| Units per wave | 502 | 251 | 100 |

Table 6: Bryan, Chowdhury, and Mobarak (2014)

| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.005 | 0.004 | 0.004 |
| expected Thompson | 0.005 | 0.005 | 0.005 |
| Thompson | 0.005 | 0.005 | 0.005 |
| non-adaptive stratified | 0.007 | 0.007 | 0.007 |
| non-adaptive | 0.009 | 0.008 | 0.007 |
| Share optimal | | | |
| modified Thompson | 0.742 | 0.770 | 0.782 |
| expected Thompson | 0.717 | 0.749 | 0.750 |
| Thompson | 0.724 | 0.749 | 0.754 |
| non-adaptive stratified | 0.666 | 0.654 | 0.668 |
| non-adaptive | 0.585 | 0.639 | 0.665 |
| Units per wave | 935 | 467 | 187 |

Table 7: Cohen, Dupas, and Schaner (2014)

| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.017 | 0.016 | 0.015 |
| expected Thompson | 0.017 | 0.016 | 0.015 |
| Thompson | 0.017 | 0.016 | 0.015 |
| non-adaptive stratified | 0.020 | 0.020 | 0.020 |
| non-adaptive | 0.026 | 0.022 | 0.020 |
| Share optimal | | | |
| modified Thompson | 0.072 | 0.076 | 0.084 |
| expected Thompson | 0.066 | 0.079 | 0.084 |
| Thompson | 0.064 | 0.077 | 0.084 |
| non-adaptive stratified | 0.051 | 0.051 | 0.053 |
| non-adaptive | 0.034 | 0.046 | 0.055 |
| Units per wave | 1080 | 540 | 216 |

**Notes:** The tables show average regret and the share of replications for which the optimal targeted treatment policy was chosen across 5,000 simulation replications. Parameters are calibrated based on the data of Ashraf et al. (2010), Bryan et al. (2014), andCohen et al. (2015). Total sample size is equal to the original sample size.

# 5 Analysis of modified Thompson sampling

In this section, we provide a discussion of the behavior of modified Thompson sampling. Our arguments draw on insights from Bubeck and Cesa-Bianchi (2012), Russo et al. (2018), and Russo (2016). We discuss how the share of units assigned to different treatments behaves in large samples, and what that implies for the probability of finding the optimal policy, expected welfare, and regret.

**Preview** We start off by motivating our modified Thompson sampling based on another algorithm. Suppose that you ran Thompson sampling on a sequence of units, but that you did not allow the same treatment to be assigned twice in a row. If the same treatment comes up twice, you keep sampling until you get a different one. As we will show, in large samples this procedure yields assignment shares which are exactly the ones that we use as assignment shares for our modified Thompson algorithm

We then turn to the large sample behavior of Thompson sampling. A fairly recent theoretical literature considers the regret behavior of various algorithms in bandit problems. There are several algorithms for which average regret goes to 0 at the optimal rate of $\log(M)/M$, as a function of total sample size $M$, including Thompson sampling. What that implies, in particular, is that after $M$ observation, only about $\log(M)$ observations have been assigned to sub-optimal treatments. This is good for the welfare of the experimental participants. But is not optimal for power, that is for the ability of a policymaker to distinguish the best option after the experiment. Thompson sampling stops learning about the performance of sub-optimal treatments too quickly, if our objective is policy choice after the experiment.

This behavior is overcome by modified Thompson sampling. By effectively forcing the algorithm to alternate between treatments, we prevent it from assigning more than half of our units to the treatment that has performed best thus far. Furthermore, this algorithm balances assignment among the sub-optimal treatments such that in large samples we have equal power for rejecting each of them. To see why this is the case, note that any sub-optimal treatment that would be assigned more often would be recognized as being sub-optimal with high probably, and thus pushed to the "back of the line" by the algorithm. Furthermore, equating power against sub-optimal treatments is optimal in terms of the rate of learning the best policy. If we were to assign one of the sub-optimal treatments more often, this would lead to a lower rate of convergence to the optimal policy.

These properties of modified Thompson sampling imply that after the experiment we end up selecting the optimal policy with higher probability than under Thompson sampling. This in turn implies that we select the optimal policy with higher probability than under non-adaptive assignment, as reflected in the simulations of Section 4.

## 5.1 Heuristic motivation for modified Thompson sampling: Alternating Thompson

Consider the following *alternating Thompson* algorithm. Assume that we sequentially assign treatments based on the Thompson probabilities $\boldsymbol{p}$, but with the twist that we don't assign the same treatment twice in a row. This algorithm defines a Markov chain for the sequence of assigned treatments. The probability of transitioning from treatment $d'$ to treatment $d \neq d'$ is given by $\frac{p^d}{1-p^{d'}}$, that is by the Thompson probability $p^d$, normalized by the probability of drawing some treatment other than $d'$. This Markov chain has a stationary distribution $\boldsymbol{q}$, where the stationary distribution satisfies the equations

$$q^d = \sum_{d' \neq d} q^{d'} \frac{p^d}{1 - p^{d'}} \tag{5.1}$$

for $d \in \{1, \ldots, k\}$. Moreover, by the mean ergodic theorem, the assignment shares of the "alternating" algorithm converge to the stationary distribution characterized by Equation (5.1). We can solve explicitly for $\boldsymbol{q}$. Denote $S = \sum_d \frac{q^d}{1-p^d}$. Then Equation 5.1 can be rewritten as

$$\frac{q^d}{p^d} = S - \frac{q^d}{1 - p^d},$$

and some algebra yields

$$\begin{aligned} q^d &= S \cdot p^d \cdot (1 - p^d), \\ S &= \sum_d p^d \cdot (1 - p^d). \end{aligned}$$

This implies that for large wave sizes the *alternating Thompson* algorithm assigns the same share of observations to each treatment as our *modified Thompson* algorithm. The latter reduces the randomness of assignments relative to alternating Thompson.

## 5.2 The large sample behavior of Thompson sampling

The idea of Thompson sampling has been around since Thompson (1933). A deeper theoretical understanding of its behavior is fairly recent, however. Bubeck and Cesa-Bianchi (2012) and chapter 8 in Russo et al. (2018) provide a review of the theoretical literature.

Key for us is the result, first shown by Agrawal and Goyal (2012) (Theorem 2), that in-sample regret for Thompson sampling (in the binary setting, with sequential arrival) satisfies the bound

$$\lim_{T\to\infty} E\left[\frac{\sum_{t=1}^{T}\Delta^{D_t}}{\log T}\right] \leq \left(\sum_{d\neq d^*}\frac{1}{(\Delta^d)^2}\right)^2,$$

where $\Delta^d = \max_{d'}\theta^{d'} - \theta^d$. As first shown by Lai and Robbins (1985), no adaptive experimental design can do better than this $\log T$ rate; the proof of this lower bound is reviewed in Section 2.3 of Bubeck and Cesa-Bianchi (2012).

What this result implies, in particular, is that after that Thompson sampling only assigns a share of units of order $\log(M)/M$ to treatments other than the optimal treatment, when there are $M$ observations total. This means that we effectively stop learning about the performance of suboptimal treatments very quickly. The posterior variance of $\theta^d$ for $d \neq d^*$ goes to zero at a rate no faster than $1/\log(M)$. Put differently, it is precisely the behavior of Thompson sampling that makes it a good choice for maximizing in-sample welfare - namely that it assigns a large share of observations to the optimal treatment - which limits its benefits for ex-post policy choice, which relies on information on sub-optimal treatment arms.

## 5.3 The large sample behavior of modified Thompson sampling

We now turn to a discussion of modified Thompson sampling. The arguments in this section build on Russo (2016), and in particular on Proposition 7, as well as Lemma 12 through 14 in Appendix G.1 therein. The following proposition characterizes the behavior of modified Thompson sampling in settings with many waves and fixed wave sizes.[4]

[4]We conjecture that similar results apply for a fixed number of waves and large wave sizes, but the proof is left for future research.

**Proposition 1** *Consider the setting of Section 2.1 with fixed wave size $N_t = N$, and the modified Thompson algorithm as defined in Section 3. As $T \to \infty$, modified Thompson assignment satisfies the following:*

1. *The share of observations assigned to the best treatment converges to* $1/2$.
2. *All the other treatments $d$ are assigned to a share of the sample which converges to a non-random share $\bar{q}^d$. $\bar{q}^d$ is such that the posterior probability of $d$ being optimal goes to* $0$ *at the same exponential rate for all sub-optimal treatments.*
3. *No other assignment algorithm for which statement 1 holds has average regret going to* $0$ *at a faster rate than modified Thompson sampling.*

The proof of this Proposition can be found in Appendix A.1. We prove these claims in several steps. First we show that each treatment is assigned infinitely often. By a basic consistency result, this implies that $p_T^d$ goes to 1 for the optimal treatment and to 0 for all other treatments. Claim 1 then follows from the definition of modified Thompson sampling. Second, we show claim 2 by contradiction. Suppose $p_t^d$ goes to 0 at a faster rate for any one of the sub-optimal treatments $d$. Then modified Thompson sampling would effectively stop assigning this treatment $d$. This in turn allows the other sub-optimal treatments to "catch up." Lastly, efficiency (claim 3) holds because the algorithm balances the rate of convergence of posterior probabilities (or equivalently, of power) across treatments. That this is optimal follows from an efficiency bound which is a corollary of a an analogous efficiency bound for best arm selection proven in Russo (2016).

Table 8: Inference for adaptive designs: A simple example

| $P$ | $Y_1$ | $Y_2$ | $\bar{Y}$ | $\hat{\theta}^B$ |
|---|---|---|---|---|
| $(1-\theta)^2$ | 0 | 0 | 0 | 1/3 |
| $\theta(1-\theta)$ | 0 | 1 | 0 | 1/3 |
| $\theta(1-\theta)$ | 1 | 0 | 1/2 | 1/2 |
| $\theta^2$ | 1 | 1 | 1 | 3/4 |
| $E[\cdot\|\theta]$ | $\theta$ | $\theta$ | $1/2 \cdot \theta(1+\theta)$ | $1/3 + \theta/6 + \theta^2/4$ |
| $E_{\theta\sim U[0,1]}[\cdot]$ | 1/2 | 1/2 | 5/12 | 1/2 |

# 6 Inference for adaptive designs

In this section we discuss inference in adaptive designs. Our discussion applies to the various adaptive algorithms considered above, including optimal assignment, Thompson sampling, expected Thompson and modified Thompson.

We will show, using a simple example, that adaptive treatment assignment biases conventional estimators, such as sample means, and affects the size of corresponding tests. We then show that Bayesian inference, by contrast, is un-affected by adaptivity. We can simply ignore adaptive assignment when calculating posterior distributions. We lastly discuss how to perform randomization tests for the null of no treatment effects. To do so, one needs to simply re-run the treatment assignment algorithm multiple times, leaving observed outcomes unchanged, to generate a randomization distribution of arbitrary test statistics.

Our discussion of inference stands somewhat outside the framework introduced in Section 2.1. In this framework, the optimal policy $d$ to choose at the end of the experiment is the one that maximizes the posterior expectation of $\theta^d$, net of costs $c^d$. Quantifications of uncertainty have no further bearing on this choice. In practice, however, experiments might have secondary purposes in addition to informing policy choice. In particular, providing a plausible range of values (for instance a confidence set) is central for academic publication of experimental results.

**A minimal example** Why does adaptivity matter for inference? Consider the following minimal example of an adaptive experiment, summarized in Table 8. Suppose that we observe a binary random variable $Y_1 \sim Ber(\theta)$. If $Y_1$ is 0, we stop the exper-

iment, if $Y_1 = 1$ we continue, and obtain another (independent) draw $Y_2 \sim Ber(\theta)$.[5] Suppose an analyst ignores the fact that the experiment had a data-dependent stopping rule, based on which we decided whether to observe $Y_2$. Would this analyst draw correct conclusions?

Consider the sample mean $\bar{Y}$, which is equal to $Y_1$ if only one draw was observed (i.e., if $Y_1 = 0$), and equal to $(Y_1 + Y_2)/2$ if two draws were observed (i.e., if $Y_1 = 1$). What is the expectation of this sample mean? There are four possible combinations of values for $(Y_1, Y_2)$, where $Y_2$ is only observed when $Y_1 = 1$. The combination $(0, 0)$, for instance, has probability $(1-\theta)^2$, and yields $\bar{Y} = Y_1 = 0$. Table 8 shows the remaining calculations which yield $E[\bar{Y}|\theta] = 1/2 \cdot \theta(1+\theta)$. In particular, whenever $\theta < 1$ we have that $\bar{Y}$ is downward-biased as an estimator of $\theta$! This example illustrates that standard frequentist inference will not be valid for adaptive designs, without modification. Such a downward bias of the sample mean similarly holds for adaptive designs such as Thompson sampling: In these designs more observations for a particular treatment are obtained, on average, whenever that treatment performed well in earlier waves.

How about Bayesian inference? As it turns out, Bayesian inference that ignores the fact that sampling was adaptive remains valid even for adaptive assignments. To illustrate, assume that we start with a uniform (i.e., $Beta(1, 1)$) prior for $\theta$. Then the posterior mean $\hat{\theta}^B$ for $\theta$ is equal to $\frac{1+Y_1}{2+1}$ if one draw of $Y_1$ is observed, and equal to $\frac{1+Y_1+Y_2}{2+2}$ if two draws are observed. This posterior mean is in fact the same whether or not we take into account that observability of $Y_2$ depends on the realization of $Y_1$. In particular, the prior mean of $\hat{\theta}^B$ is indeed equal to $1/2$, the prior mean of $\theta$. This contrasts with the prior mean of $\bar{Y}$, which is equal to $5/12$.

**General validity of standard Bayesian inference** The validity of standard Bayesian inference is not specific to this example. Bayesian inference that ignores adaptivity remains in fact correct in the context of any adaptive experimental design setting. This follows from the following simple proposition.

**Assumption 1** *Consider a set of units $i = 1, 2, \ldots, M$ with potential outcomes $(Y_i^1, \ldots, Y_i^k)$. Denote the likelihood of $Y_i^d$, indexed by the parameter vector $\boldsymbol{\theta}$, by $f_{\boldsymbol{\theta}}^d(Y_i^d)$.*
*Let $H_i = (D_1, \ldots, D_{i-1}, Y_1, \ldots, Y_{i-1})$ be the history of treatments and outcomes before*

[5] This is a stylized version of the behavior of the algorithms considered above, which sample more units for treatment arms that had better performance in the past.

*i. Suppose that the treatment assigned to unit i is a function of $H_i$, as well as possibly a randomization device $U_i$.*[6]

**Proposition 2** *Under Assumption 1, the likelihood of $(D_1, \ldots, D_M, Y_1, \ldots, Y_M)$ equals*

$$\prod_{i=1}^{M} f_{\boldsymbol{\theta}}^{D_i}(Y_i),$$

*up to a constant that does not depend on $\boldsymbol{\theta}$.*

**Proof:** This follows immediately from the factorization

$$P(D_1, \ldots, D_M, Y_1, \ldots, Y_M | \boldsymbol{\theta}) = \prod_{i=1}^{M} P(D_i | H_i, \boldsymbol{\theta}) \cdot P(Y_i | D_i, H_i, \boldsymbol{\theta}),$$

where $P(D_i|H_i, \boldsymbol{\theta})$ does not depend on $\theta$, and $P(Y_i|D_i, H_i, \boldsymbol{\theta}) = f_{\boldsymbol{\theta}}^{D_i}(Y_i)$. □

**Construction of valid randomization tests** A popular method for inference in randomized experiments are randomization tests (also known as permutation tests). Consider again the setting of Assumption 1, and the following null hypothesis.

**H0 1** *For each unit i, the potential outcomes are the same for all treatments,*

$$Y_i^d = Y_i \quad \forall\, i \,\forall\, d.$$

A randomization test in this setting can be constructed by simulating treatment vectors $\tilde{D}_i$ using a scheme that iterates from $i = 1 \ldots M$, drawing $\tilde{U}_i$ from the same distribution as $U_i$, defining

$$\tilde{H}_i = (\tilde{D}_1, \ldots, \tilde{D}_{i-1}, Y_1, \ldots, Y_{i-1}),$$

and setting

$$\tilde{D}_i = d_i(\tilde{H}_i, \tilde{U}_i).$$

The following proposition is immediate.

[6] This setting includes as a special case treatment assignment in waves, and binary outcomes, as in Section 2.1, where treatment assignments only depend on the outcomes of previous waves, and where $f_{\boldsymbol{\theta}}^d(y) = y \cdot \theta^d + (1 - y) \cdot (1 - \theta^d)$.

**Proposition 3** *Consider the setting of Assumption 1.*

1. *Under the null hypothesis 1, $\tilde{H}_M$ has the same distribution as $H_M$, conditional on $(Y_1, \ldots, Y_M)$.*

2. *Let $T(H_M)$ be any test statistic, and let $T^\alpha$ be the $(1-\alpha)$ quantile of the distribution of $T(\tilde{H}_M)$ over draws of $\tilde{U}_1, \ldots, \tilde{U}_M$. Then a test which rejects iff $T(H_M) > T^\alpha$ controls size at level $\alpha$.*

# 7 Implementation in the field

# 8 Conclusion

Experiments can have different goals. The goal might be to get precise estimates of treatment effects. This goal rationalizes conventional experimental design recommendations. The goal might alternatively be to maximize the outcomes of experimental participants, through adaptive assignment. This goal rationalizes algorithms from the multi-armed Bandit literature, including Thompson sampling. Such algorithms trade off "exploitation" (maximizing the outcomes of current participants) and "exploration" (learning to achieve better assignments for future participants).

Yet another goal, and the one that we consider in this paper, is to use the experiment to choose a policy for large scale implementation after conclusion of the experiment. We first consider fully optimal experimental designs, maximizing expected welfare. These designs shift treatment assignment toward the better performing arms over time, despite the lack of an "exploitation" motive. The optimal designs are given by the solution to a dynamic stochastic optimization problem and can be derived using backward induction.

The computational costs for optimization grow quickly for realistic sample sizes, however. This motivates the use of simpler algorithms. We propose one such algorithm, modified Thompson sampling, which approximates the behavior of optimal designs. As we prove, modified Thompson sampling shifts assignment toward the better performing treatments, but makes sure that no more than half the sample is assigned to the best treatment. It furthermore balances assignment to the other treatments such that the posterior probability of each of them being optimal goes to 0 at the same rate. This behavior is exactly what is required for optimal convergence rates of expected welfare.

Simulations calibrated to parameters from actual field experiments in development economics confirm these conclusions. We find that modified Thompson sampling generates higher expected welfare and a higher probability of picking the optimal policy, relative to both conventional assignment and Thompson sampling. More generally, we find that the distribution of realized welfare across repeated simulations first order stochastically dominates the distribution under alternative algorithms.

All of these conclusions carry over to settings with covariates and targeted treatment assignment policies, where we consider hierarchical Bayesian models. Inference for adaptive experimental designs needs to take into account adaptivity. We discuss both Bayesian inference and randomization inference.

# References


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

Ashraf, N., Berry, J., and Shapiro, J. M. (2010). Can higher prices stimulate product use? Evidence from a field experiment in Zambia. *American Economic Review*, 100(5):2383–2413.

Athey, S. and Imbens, G. W. (2017). The econometrics of randomized experimentsa. In *Handbook of Economic Field Experiments*, volume 1, pages 73–140. Elsevier.

Banerjee, A., Duflo, E., and Kremer, M. (2016). The influence of randomized controlled trials on development economics research and on development policy. *Mimeo MIT*.

Berry, D. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery*, 5(1):27–36.

Bryan, G., Chowdhury, S., and Mobarak, A. M. (2014). Underinvestment in a profitable technology: The case of seasonal migration in Bangladesh. *Econometrica*, 82(5):1671–1748.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

Cohen, J., Dupas, P., and Schaner, S. (2015). Price subsidies, diagnostic tests, and targeting of malaria treatment: evidence from a randomized controlled trial. *American Economic Review*, 105(2):609–45.

Conde-Agudelo, A., Belizán, J. M., and Diaz-Rossello, J. (2012). Cochrane review: Kangaroo mother care to reduce morbidity and mortality in low birthweight infants. *Evidence-Based Child Health: A Cochrane Review Journal*, 7(2):760–876.

Duflo, E. (2017). Richard T. Ely lecture: The economist as plumber. *American Economic Review*, 107(5):1–26.

Duflo, E. and Banerjee, A., editors (2017). *Handbook of Field Experiments*, volume 1. Elsevier.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Taylor & Francis.

Ghavamzadeh, M., Mannor, S., Pineau, J., and Tamar, A. (2015). Bayesian reinforcement learning: A survey. *Foundations and TrendsÂ® in Machine Learning*, 8(5-6):359–483.

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

Russo, D. (2016). Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418.

Russo, D. J., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. (2018). A Tutorial on Thompson Sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Weber, R. et al. (1992). On the Gittins index for multiarmed bandits. *The Annals of Applied Probability*, 2(4):1024–1033.

# A Supplementary Appendix

A.1 Proofs . . . . . 43
A.2 Additional simulation results . . . . . 45
A.2.1 No covariates . . . . . 45
A.2.2 Targeting based on covariates . . . . . 52
A.3 Additional optimal design example plots . . . . . 57

## A.1 Proofs

**Proof of Proposition 1:**
Recall that under modified Thompson sampling, a share

$$q_t^d = \frac{p_t^d \cdot (1 - p_t^d)}{\sum_d p_t^{d'} \cdot (1 - p_t^{d'})}$$

of wave $t$ is assigned to treatment $d$, where $p_t^d$ is the posterior probability that $d$ is optimal.

1. **Each treatment is assigned infinitely often.**
   Suppose otherwise. Then there is some treatment which is only assigned a finite number of times, and is not assigned anymore after some wave $t'$, so that $q_t^d = 0$ for $t > t'$. The posterior probability $p_t^d$ of this treatment being optimal is bounded away from 0 for $t > t'$ by Lemma 14 in Russo (2016).

   Note now that under modified Thompson sampling, the denominator in the expression defining $q_t^d$ is bounded above by 1, and thus the probability $q_t^d$ of being assigned to treatment $d$ is bounded below by $p_t^d \cdot (1 - p_t^d)$. It follows that $q_t^d$ is bounded away from 0 when the same holds for $p_t^d$. The claim follows by contradiction.

2. **The share of observations assigned to the best treatment converges to 1/2 as $T \to \infty$.**
   Since each treatment is assigned infinitely often, we have that $p_t^d \to 1$ for the optimal treatment $d$, again by Lemma 14 in Russo (2016).

   Concavity of the denominator of the expression defining $q_t^d$, as a function of the vector $\boldsymbol{p}_t$, implies

$$q_t^d \leq \frac{1}{2}$$

(equality is achieved when $p_t^{d'}$ is equal to 0 for all but two values of $d'$) and

$$q_t^d \geq \frac{p_t^d}{p_t^d + 1 - (1 - p_t^d)/(k-1)}$$

(equality is achieved when $p_t^{d'}$ is equal to $\frac{1-p_t^d}{k-1}$ for all $d' \neq d$), where

$$\frac{p_t^d}{p_t^d + 1 - (1 - p_t^d)/(k-1)} \geq \frac{p_t^d}{p_t^d + 1} \to \frac{1}{2}$$

The claim follows.

3. **All the other treatments $d$ are assigned to a share of the sample which converges to a non-random share $\bar{q}^d$. $\bar{q}^d$ is such that the posterior probability of $d$ being optimal goes to $0$ at the same exponential rate for all sub-optimal treatments.**
Consider two treatments $d, d'$. By definition of $q_t^d$,

$$q_t^d \leq \frac{p_t^d \cdot (1 - p_t^d)}{p_t^{d'} \cdot (1 - p_t^{d'})} \leq 4 \frac{p_t^d}{p_t^{d'}},$$

where the second inequality holds as long as $p_t^{d'} \leq 1/2$. Lemma 13 in Russo (2016) then implies that $q_t^d$ converges to 0 at an exponential rate, along any subsequence of $t$ for which the share of observations assigned to $d$ exceeds the share $\bar{q}^d$.

By Lemma 12 in Russo (2016), this in turn implies that the share of observations assigned to $d$ has to converge to $\bar{q}^d$.

4. **No other assignment algorithm for which statement 1 holds has average regret going to $0$ at a faster rate than modified Thompson sampling.**
This is an immediate corollary of Proposition 7 Russo (2016), once we note that the rate of convergence of average regret to 0 is the same as the rate of convergence of the probability of choosing a sub-optimal treatment.

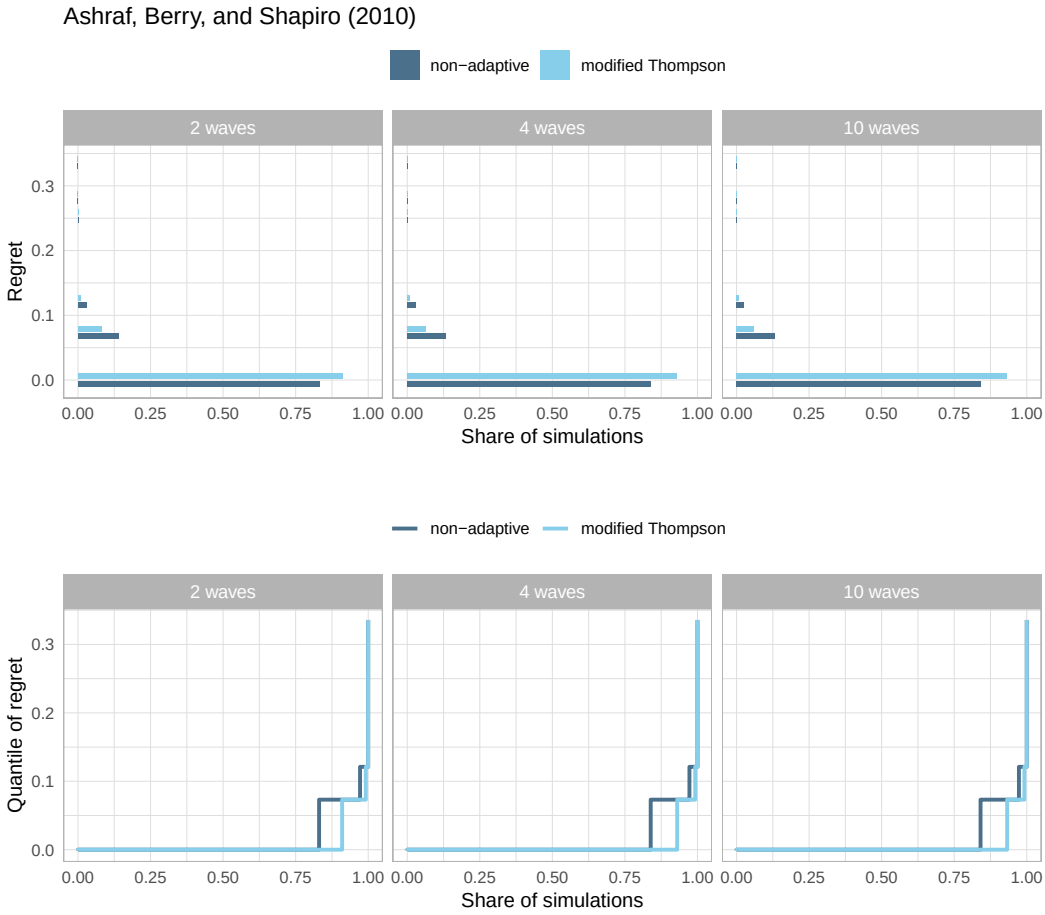
## A.2 Additional simulation results

In this section, we present additional simulation results based on the same calibrated parameter values as in Section 4. We first consider the case of no covariates, and then present results with targeting based on covariates. For each of these, we first present simulation results based on half the original sample size, and then simulation results based on 1.5 times the original sample size.

### A.2.1 No covariates

Table A1: Ashraf, Berry, and Shapiro (2010)

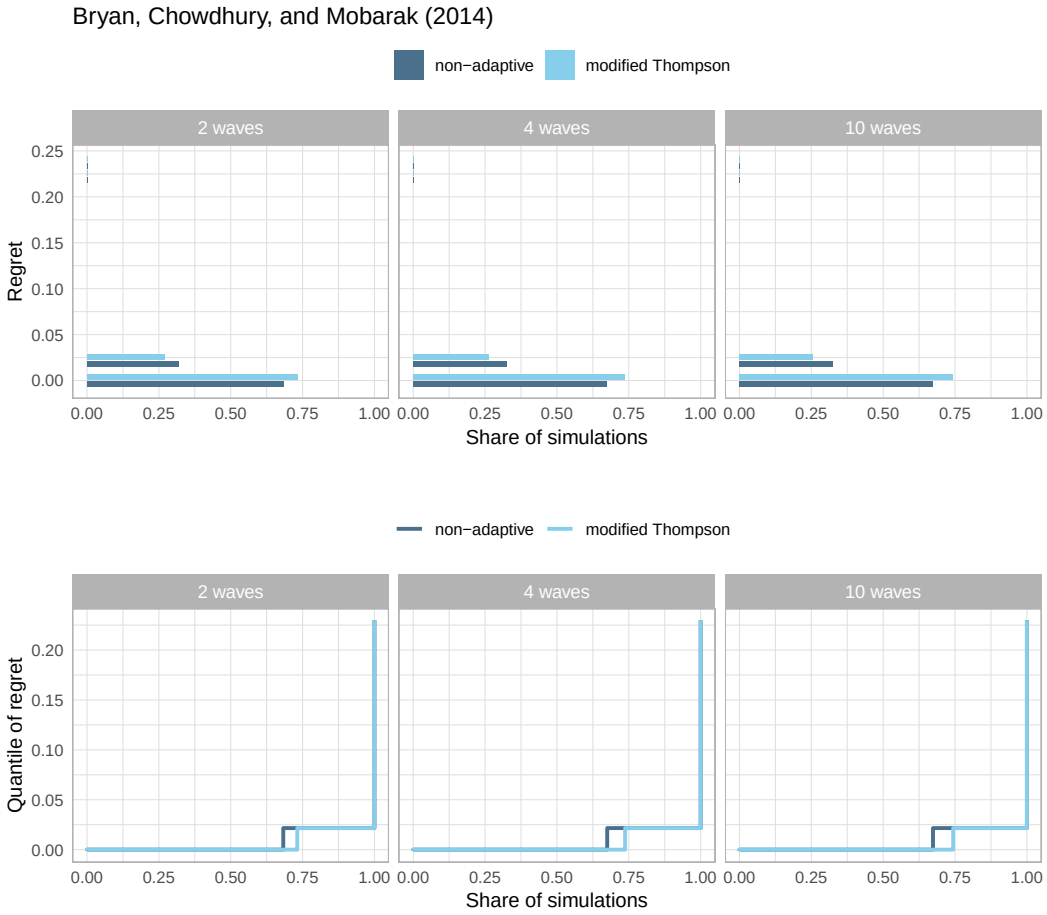
| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.007 | 0.005 | 0.005 |
| expected Thompson | 0.008 | 0.007 | 0.006 |
| Thompson | 0.008 | 0.007 | 0.006 |
| non-adaptive | 0.014 | 0.013 | 0.013 |
| Share optimal | | | |
| modified Thompson | 0.910 | 0.930 | 0.932 |
| expected Thompson | 0.898 | 0.914 | 0.928 |
| Thompson | 0.898 | 0.917 | 0.918 |
| non-adaptive | 0.831 | 0.839 | 0.841 |
| Units per wave | 251 | 125 | 50 |




**Notes:** The table shows average regret and the share of replications for which the optimal treatment was chosen across 20,000 simulation replications. The histograms and quantiles show the distribution of regret. Parameters are calibrated based on the data of Ashraf et al. (2010). Total sample size is equal to half the original.

Table A2: Bryan, Chowdhury, and Mobarak (2014)

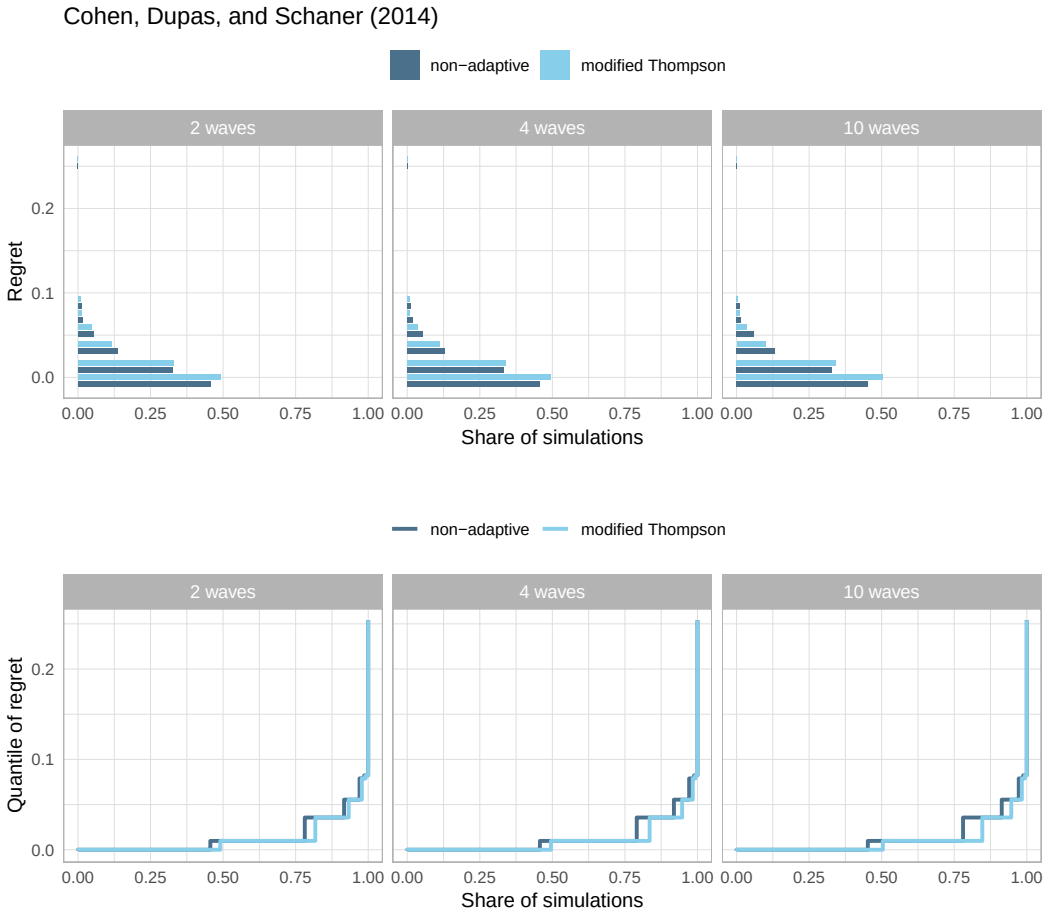
| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.006 | 0.006 | 0.006 |
| expected Thompson | 0.006 | 0.006 | 0.006 |
| Thompson | 0.006 | 0.006 | 0.006 |
| non-adaptive | 0.007 | 0.007 | 0.007 |
| Share optimal | | | |
| modified Thompson | 0.732 | 0.737 | 0.744 |
| expected Thompson | 0.706 | 0.719 | 0.725 |
| Thompson | 0.710 | 0.718 | 0.725 |
| non-adaptive | 0.683 | 0.675 | 0.673 |
| Units per wave | 467 | 233 | 93 |




**Notes:** The table shows average regret and the share of replications for which the optimal treatment was chosen across 20,000 simulation replications. The histograms and quantiles show the distribution of regret. Parameters are calibrated based on the data of Bryan et al. (2014). Total sample size is equal to half the original.

Table A3: Cohen, Dupas, and Schaner (2014)

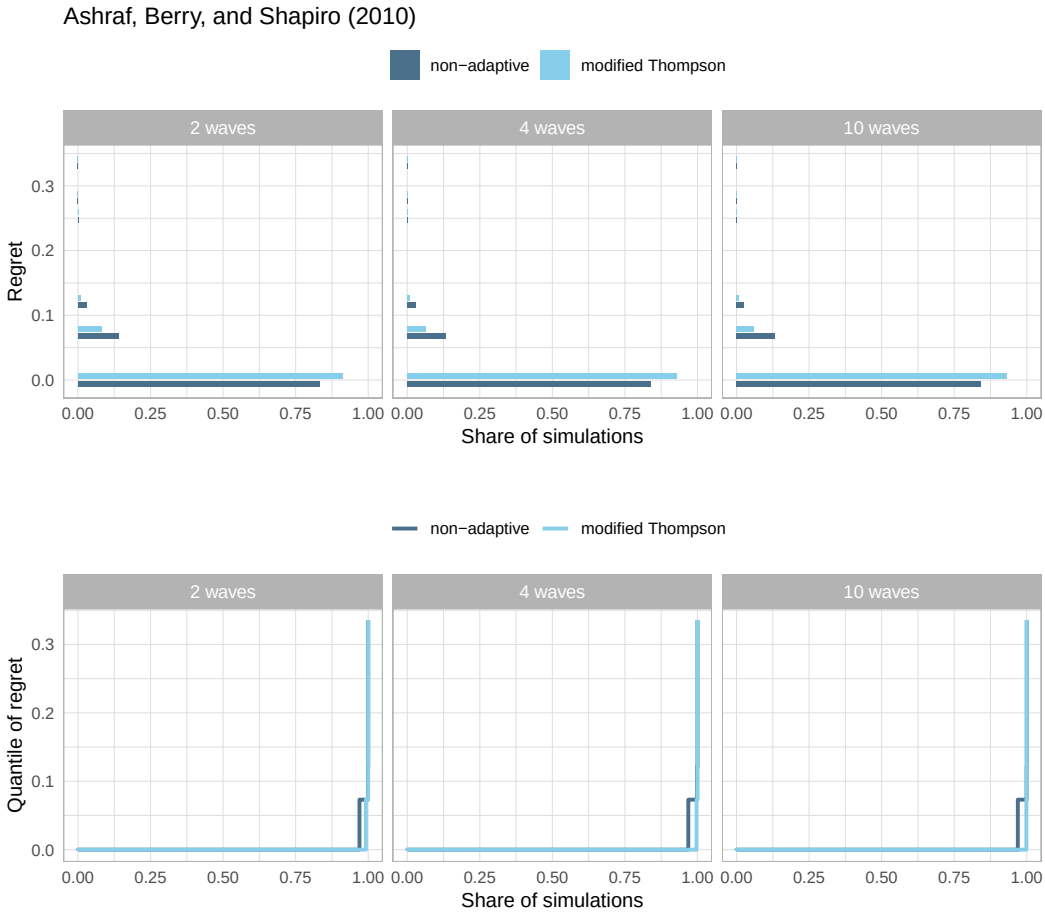
| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.012 | 0.011 | 0.010 |
| expected Thompson | 0.012 | 0.011 | 0.010 |
| Thompson | 0.012 | 0.011 | 0.010 |
| non-adaptive | 0.013 | 0.013 | 0.013 |
| Share optimal | | | |
| modified Thompson | 0.490 | 0.495 | 0.504 |
| expected Thompson | 0.479 | 0.491 | 0.494 |
| Thompson | 0.484 | 0.501 | 0.500 |
| non-adaptive | 0.456 | 0.457 | 0.452 |
| Units per wave | 540 | 270 | 108 |




**Notes:** The table shows average regret and the share of replications for which the optimal treatment was chosen across 20,000 simulation replications. The histograms and quantiles show the distribution of regret. Parameters are calibrated based on the data of Cohen et al. (2015). Total sample size is equal to half the original.

Table A4: Ashraf, Berry, and Shapiro (2010)

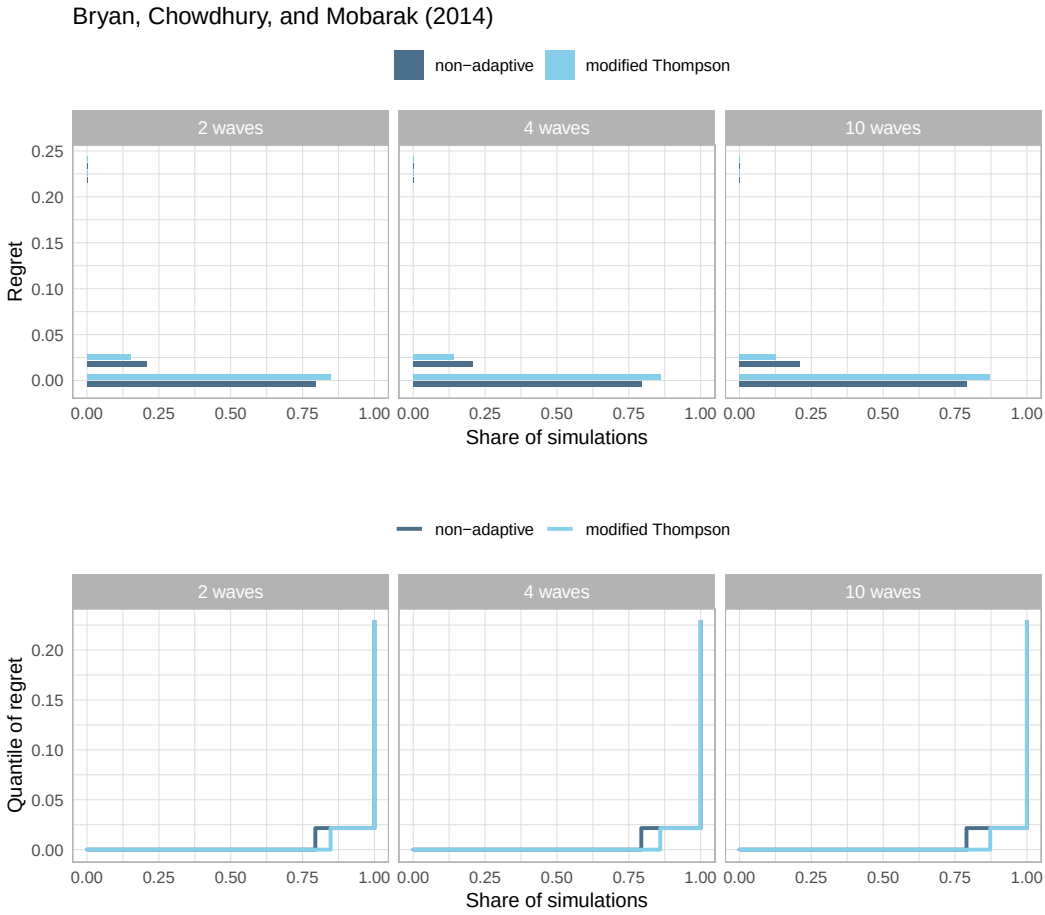
| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.001 | 0.000 | 0.000 |
| expected Thompson | 0.001 | 0.000 | 0.000 |
| Thompson | 0.001 | 0.000 | 0.000 |
| non-adaptive | 0.002 | 0.002 | 0.002 |
| Share optimal | | | |
| modified Thompson | 0.992 | 0.997 | 0.999 |
| expected Thompson | 0.991 | 0.996 | 0.995 |
| Thompson | 0.991 | 0.994 | 0.995 |
| non-adaptive | 0.970 | 0.968 | 0.969 |
| Units per wave | 753 | 376 | 150 |




**Notes:** The table shows average regret and the share of replications for which the optimal treatment was chosen across 20,000 simulation replications. The histograms and quantiles show the distribution of regret. Parameters are calibrated based on the data of Ashraf et al. (2010). Total sample size is equal to 1.5 times the original.

Table A5: Bryan, Chowdhury, and Mobarak (2014)

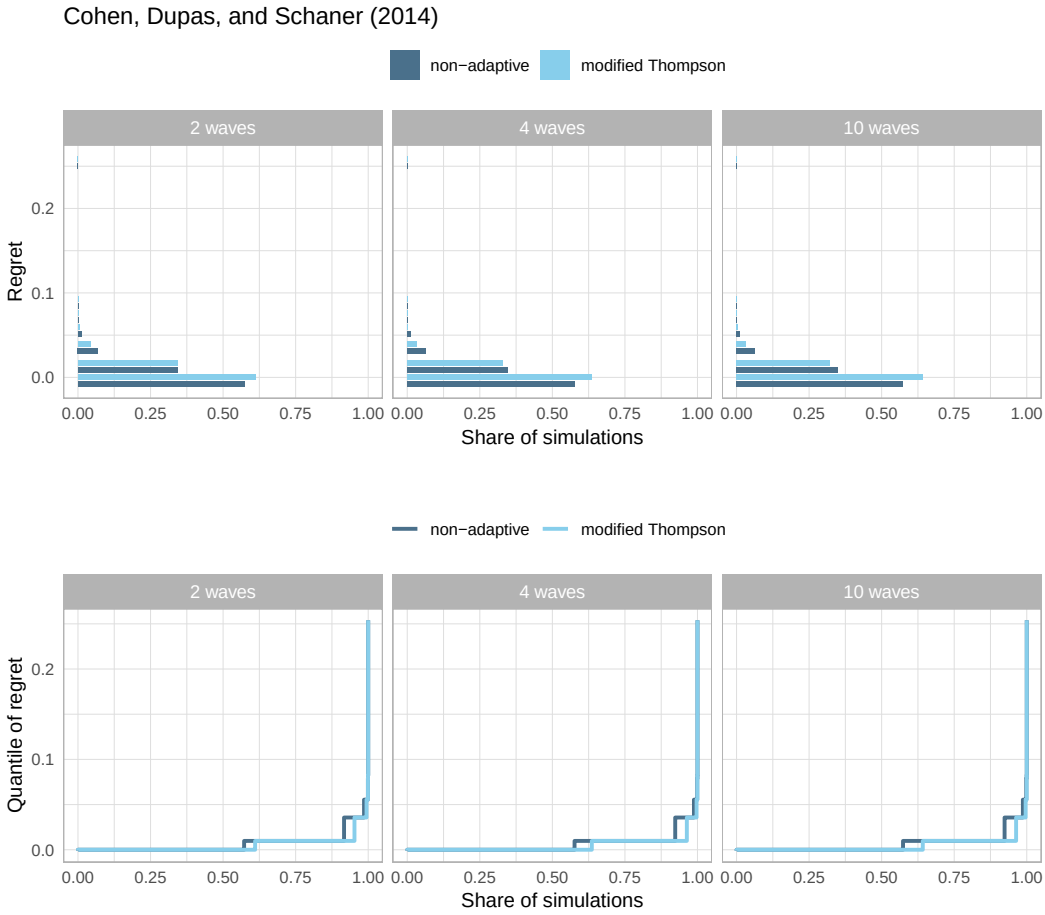
| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.003 | 0.003 | 0.003 |
| expected Thompson | 0.004 | 0.003 | 0.003 |
| Thompson | 0.004 | 0.003 | 0.003 |
| non-adaptive | 0.004 | 0.004 | 0.005 |
| Share optimal | | | |
| modified Thompson | 0.847 | 0.859 | 0.872 |
| expected Thompson | 0.825 | 0.841 | 0.854 |
| Thompson | 0.821 | 0.850 | 0.855 |
| non-adaptive | 0.794 | 0.793 | 0.790 |
| Units per wave | 1402 | 701 | 280 |




**Notes:** The table shows average regret and the share of replications for which the optimal treatment was chosen across 20,000 simulation replications. The histograms and quantiles show the distribution of regret. Parameters are calibrated based on the data of Bryan et al. (2014). Total sample size is equal to 1.5 times the original.

Table A6: Cohen, Dupas, and Schaner (2014)

| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.005 | 0.005 | 0.005 |
| expected Thompson | 0.006 | 0.005 | 0.005 |
| Thompson | 0.005 | 0.005 | 0.004 |
| non-adaptive | 0.007 | 0.006 | 0.006 |
| Share optimal | | | |
| modified Thompson | 0.610 | 0.636 | 0.642 |
| expected Thompson | 0.613 | 0.633 | 0.643 |
| Thompson | 0.608 | 0.628 | 0.649 |
| non-adaptive | 0.573 | 0.576 | 0.574 |
| Units per wave | 1620 | 810 | 324 |




**Notes:** The table shows average regret and the share of replications for which the optimal treatment was chosen across 20,000 simulation replications. The histograms and quantiles show the distribution of regret. Parameters are calibrated based on the data of Cohen et al. (2015). Total sample size is equal to 1.5 times the original.

### A.2.2 Targeting based on covariates

**Half of the original sample size**

Table A7: Ashraf, Berry, and Shapiro (2010)

| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.020 | 0.018 | 0.018 |
| expected Thompson | 0.021 | 0.019 | 0.018 |
| Thompson | 0.021 | 0.019 | 0.018 |
| non-adaptive stratified | 0.026 | 0.026 | 0.027 |
| non-adaptive | 0.035 | 0.028 | 0.027 |
| Share optimal | | | |
| modified Thompson | 0.122 | 0.138 | 0.117 |
| expected Thompson | 0.101 | 0.118 | 0.122 |
| Thompson | 0.108 | 0.114 | 0.122 |
| non-adaptive stratified | 0.088 | 0.074 | 0.072 |
| non-adaptive | 0.060 | 0.076 | 0.076 |
| Units per wave | 251 | 125 | 50 |

Table A8: Bryan, Chowdhury, and Mobarak (2014)

| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.009 | 0.008 | 0.007 |
| expected Thompson | 0.009 | 0.009 | 0.008 |
| Thompson | 0.009 | 0.009 | 0.008 |
| non-adaptive stratified | 0.011 | 0.011 | 0.011 |
| non-adaptive | 0.013 | 0.011 | 0.011 |
| Share optimal | | | |
| modified Thompson | 0.601 | 0.623 | 0.638 |
| expected Thompson | 0.567 | 0.592 | 0.595 |
| Thompson | 0.591 | 0.588 | 0.603 |
| non-adaptive stratified | 0.520 | 0.533 | 0.514 |
| non-adaptive | 0.452 | 0.513 | 0.516 |
| Units per wave | 467 | 233 | 93 |

Table A9: Cohen, Dupas, and Schaner (2014)

| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.024 | 0.024 | 0.023 |
| expected Thompson | 0.026 | 0.024 | 0.023 |
| Thompson | 0.025 | 0.023 | 0.023 |
| non-adaptive stratified | 0.029 | 0.029 | 0.028 |
| non-adaptive | 0.036 | 0.030 | 0.030 |
| Share optimal | | | |
| modified Thompson | 0.037 | 0.038 | 0.047 |
| expected Thompson | 0.027 | 0.040 | 0.041 |
| Thompson | 0.033 | 0.040 | 0.041 |
| non-adaptive stratified | 0.026 | 0.025 | 0.025 |
| non-adaptive | 0.016 | 0.017 | 0.022 |
| Units per wave | 540 | 270 | 108 |

**Notes:** The tables show average regret and the share of replications for which the optimal targeted treatment policy was chosen across 5,000 simulation replications. Parameters are calibrated based on the data of Ashraf et al. (2010), Bryan et al. (2014), andCohen et al. (2015). Total sample size is equal to half the original.

**1.5 times the original sample size**

Table A10: Ashraf, Berry, and Shapiro (2010)

| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.009 | 0.008 | 0.008 |
| expected Thompson | 0.009 | 0.008 | 0.008 |
| Thompson | 0.009 | 0.008 | 0.008 |
| non-adaptive stratified | 0.012 | 0.012 | 0.012 |
| non-adaptive | 0.018 | 0.014 | 0.013 |
| Share optimal | | | |
| modified Thompson | 0.242 | 0.263 | 0.265 |
| expected Thompson | 0.231 | 0.261 | 0.272 |
| Thompson | 0.236 | 0.254 | 0.264 |
| non-adaptive stratified | 0.184 | 0.182 | 0.188 |
| non-adaptive | 0.130 | 0.164 | 0.179 |
| Units per wave | 753 | 376 | 150 |

Table A11: Bryan, Chowdhury, and Mobarak (2014)

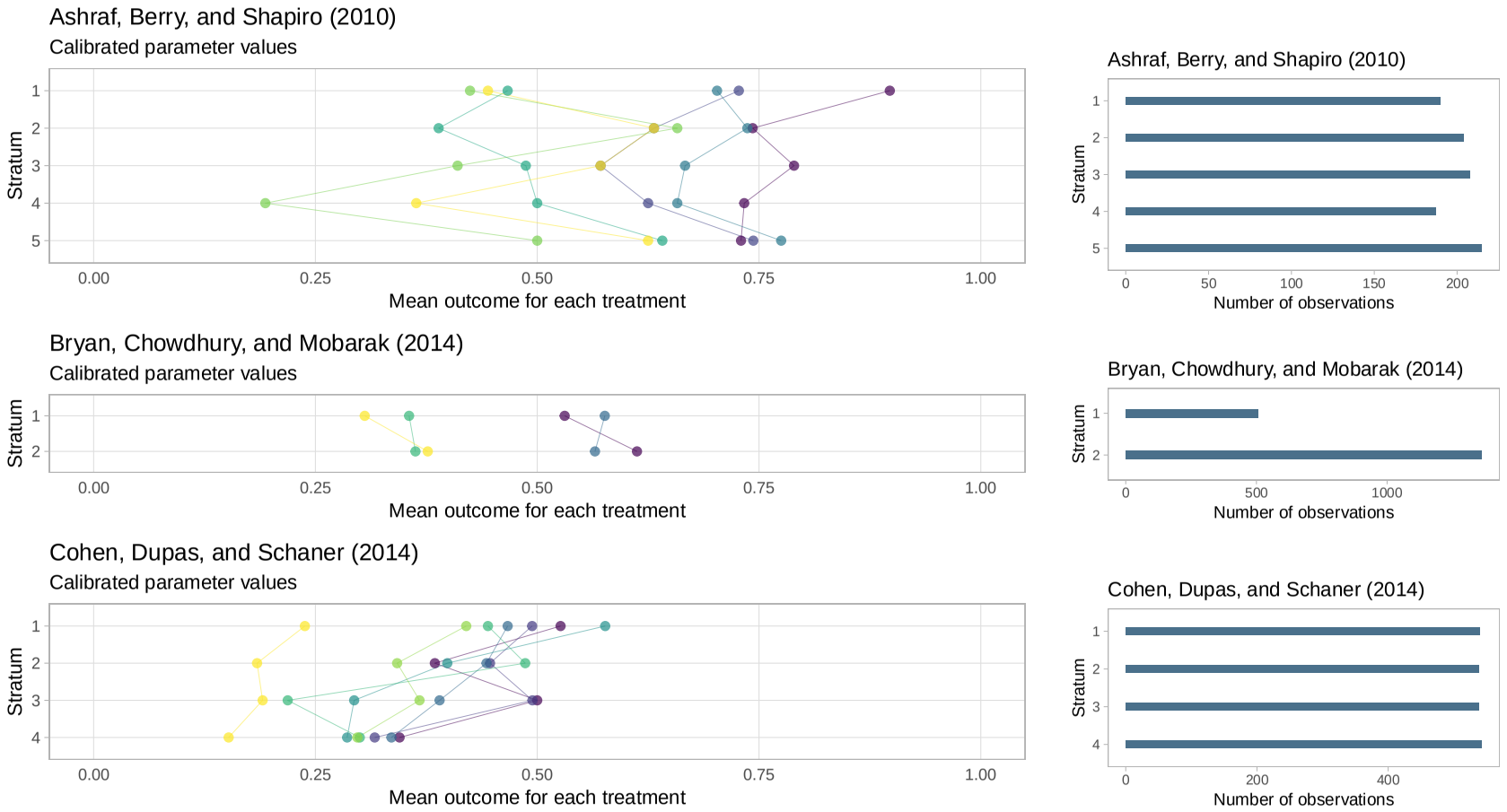
| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.003 | 0.002 | 0.002 |
| expected Thompson | 0.003 | 0.003 | 0.003 |
| Thompson | 0.003 | 0.003 | 0.003 |
| non-adaptive stratified | 0.005 | 0.005 | 0.004 |
| non-adaptive | 0.007 | 0.005 | 0.005 |
| Share optimal | | | |
| modified Thompson | 0.829 | 0.852 | 0.861 |
| expected Thompson | 0.805 | 0.829 | 0.827 |
| Thompson | 0.802 | 0.823 | 0.830 |
| non-adaptive stratified | 0.750 | 0.741 | 0.762 |
| non-adaptive | 0.659 | 0.727 | 0.747 |
| Units per wave | 1402 | 701 | 280 |

Table A12: Cohen, Dupas, and Schaner (2014)

| Statistic | 2 waves | 4 waves | 10 waves |
|---|---|---|---|
| Regret | | | |
| modified Thompson | 0.012 | 0.012 | 0.011 |
| expected Thompson | 0.013 | 0.012 | 0.011 |
| Thompson | 0.013 | 0.012 | 0.011 |
| non-adaptive stratified | 0.015 | 0.015 | 0.015 |
| non-adaptive | 0.021 | 0.017 | 0.016 |
| Share optimal | | | |
| modified Thompson | 0.110 | 0.112 | 0.124 |
| expected Thompson | 0.095 | 0.114 | 0.116 |
| Thompson | 0.101 | 0.109 | 0.119 |
| non-adaptive stratified | 0.075 | 0.079 | 0.077 |
| non-adaptive | 0.049 | 0.069 | 0.076 |
| Units per wave | 1620 | 810 | 324 |

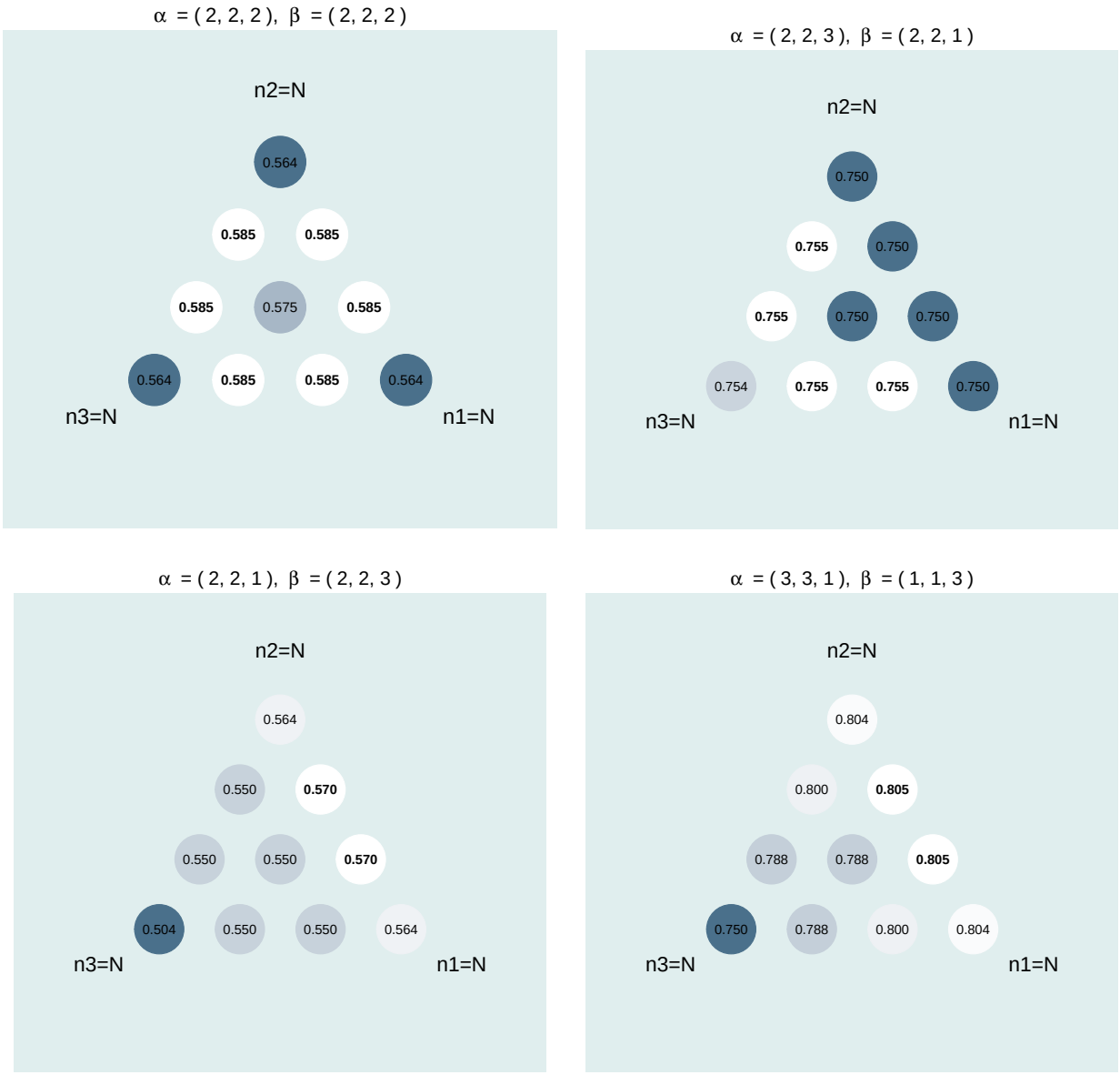
**Notes:** The tables show average regret and the share of replications for which the optimal targeted treatment policy was chosen across 5,000 simulation replications. Parameters are calibrated based on the data of Ashraf et al. (2010), Bryan et al. (2014), andCohen et al. (2015). Total sample size is equal to 1.5 times the original.

Figure A1: Average outcomes across treatment arms and strata in published experiments



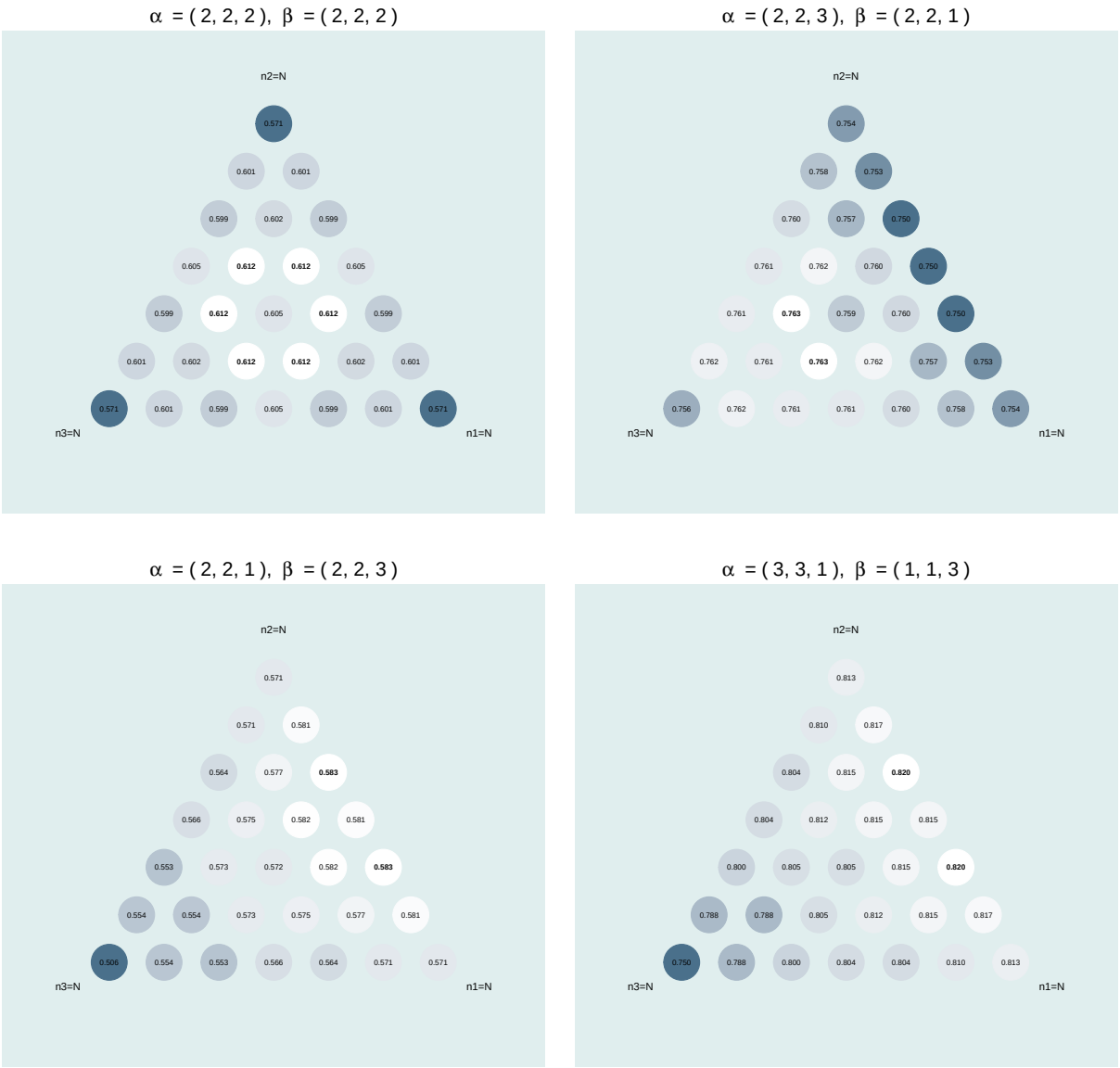
## A.3 Additional optimal design example plots

Figure A2: Expected welfare as a function of treatment assignment, sample size 3



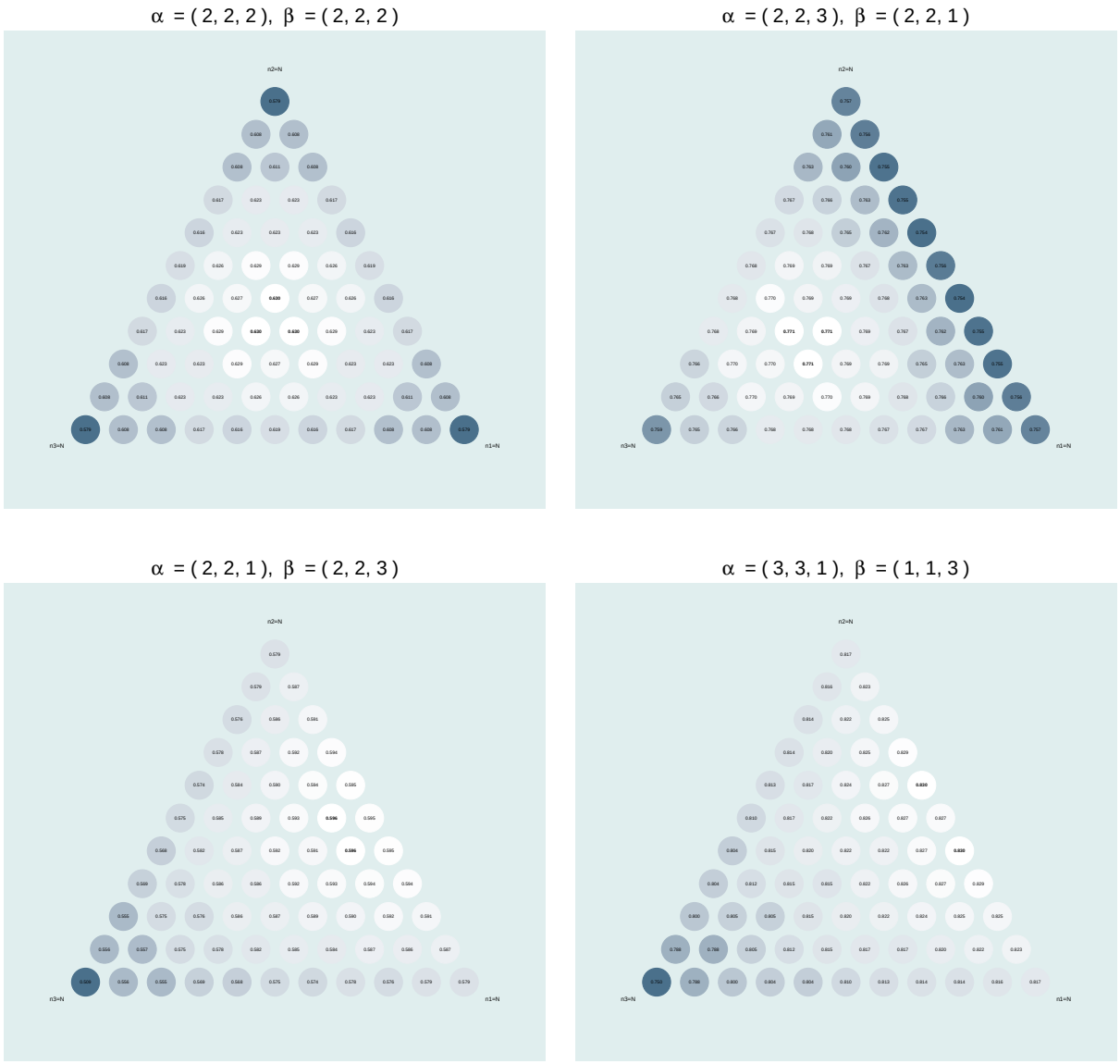

**Notes:** This figure shows the expected welfare $U_2$ as a function of treatment assignment $\boldsymbol{n}_2$ in wave 2 (size 3), taking as given the *Beta*-prior parameters $\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1$ determined by the outcomes of wave 1 (size 6). Note that the color scaling differs across figures for better readability.

Figure A3: Expected welfare as a function of treatment assignment, sample size 6



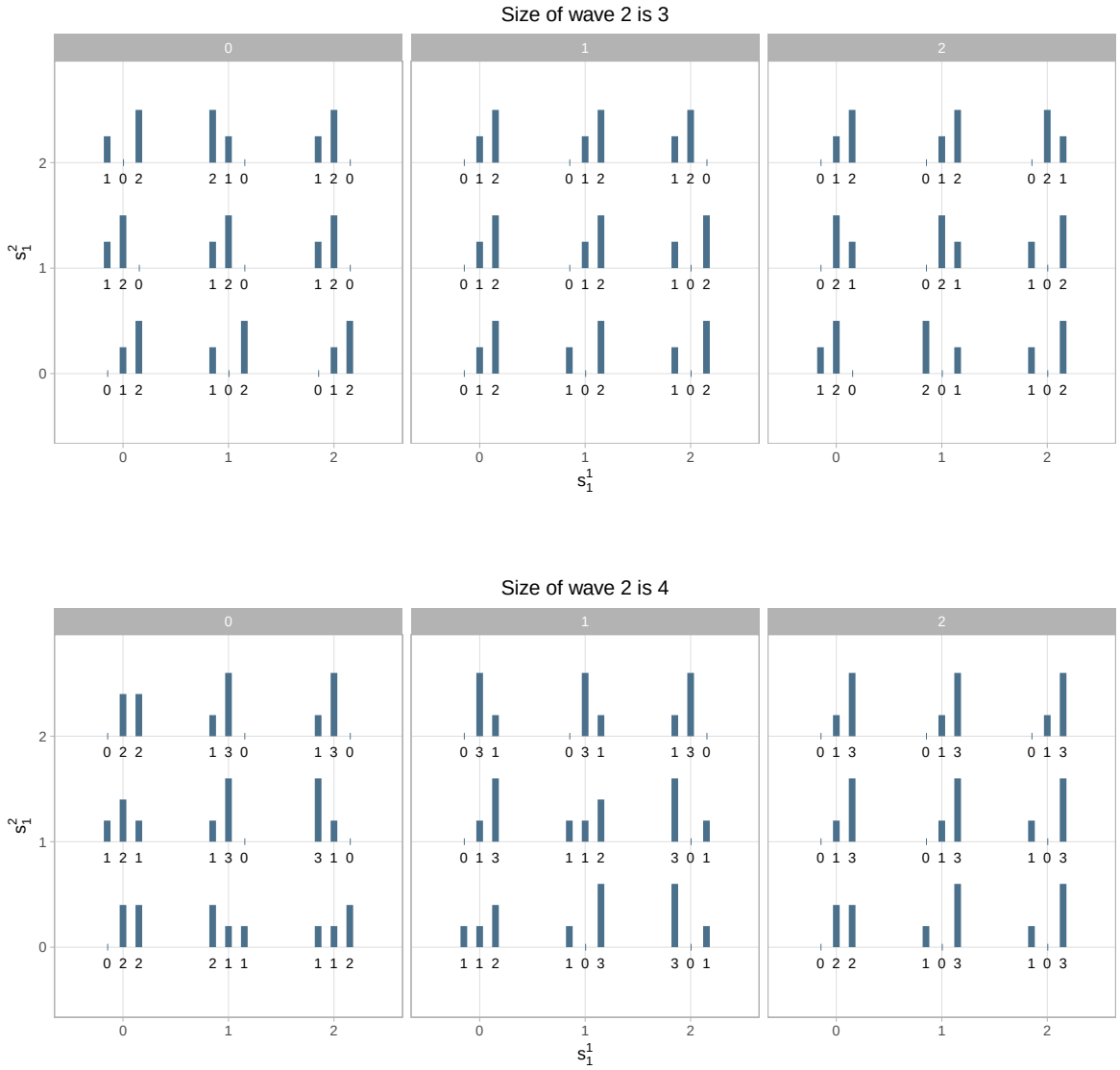

**Notes:** This figure shows the expected welfare $U_2$ as a function of treatment assignment $\boldsymbol{n}_2$ in wave 2 (size 6), taking as given the *Beta*-prior parameters $\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1$ determined by the outcomes of wave 1 (size 6). Note that the color scaling differs across figures for better readability.

Figure A4: Expected welfare as a function of treatment assignment, sample size 10



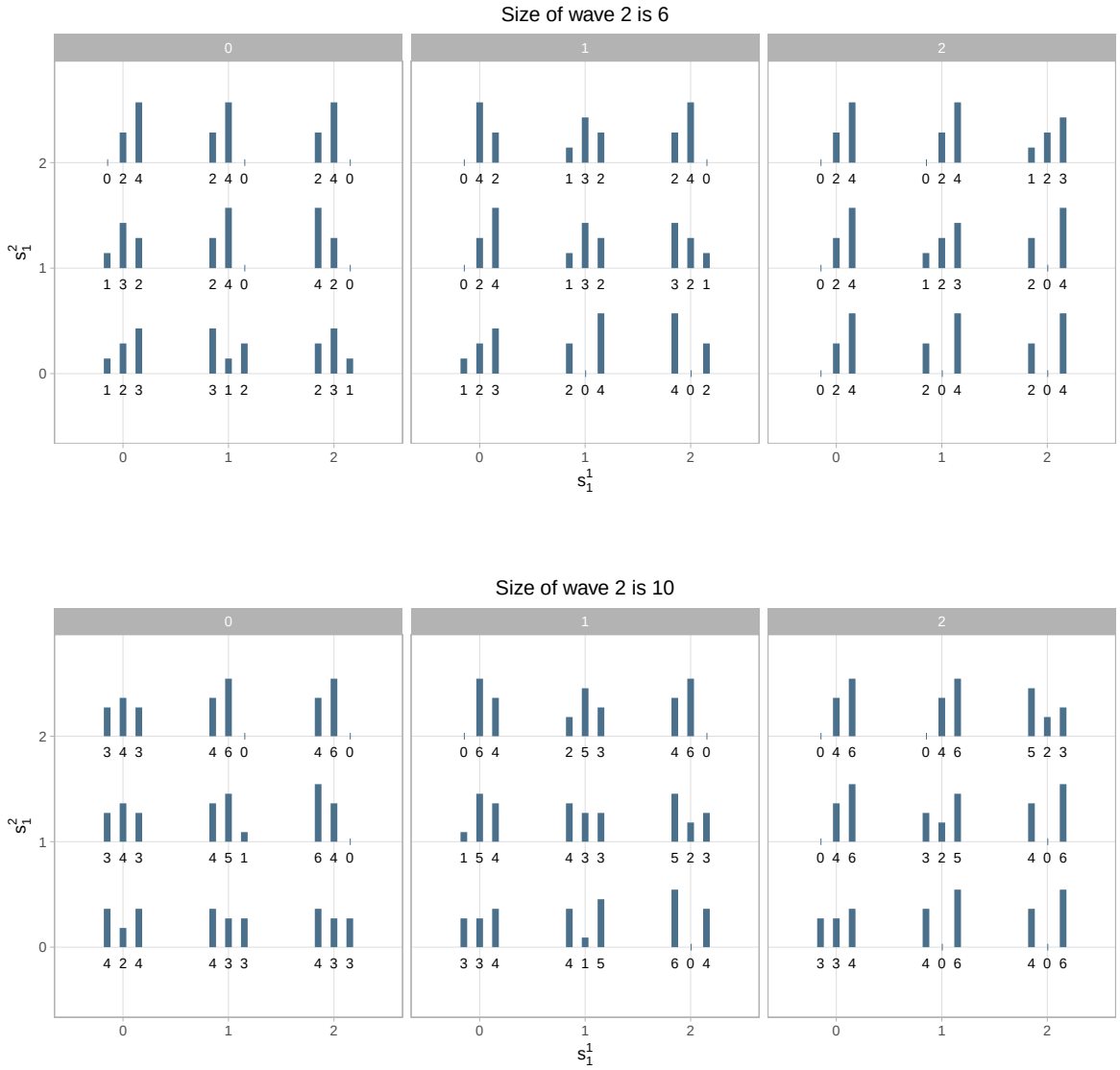

**Notes:** This figure shows the expected welfare $U_2$ as a function of treatment assignment $\boldsymbol{n}_2$ in wave 2 (size 10), taking as given the *Beta*-prior parameters $\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1$ determined by the outcomes of wave 1 (size 6). Note that the color scaling differs across figures for better readability.

Figure A5: Optimal treatment assignments in wave 2




**Notes:** These figures shows optimal treatment assignments in wave 2 as a function of the number of successes in wave 1 for treatment 1 (horizontal axis), treatment 2 (vertical axis), and treatment 3 (panels). These plots correspond to the settings of Figure A2 and 2.2. Ties between different optimal assignments are broken arbitrarily.

Figure A6: Optimal treatment assignments in wave 2




**Notes:** These figures shows optimal treatment assignments in wave 2 as a function of the number of successes in wave 1 for treatment 1 (horizontal axis), treatment 2 (vertical axis), and treatment 3 (panels). These plots correspond to the settings of Figure A3 and A4. Ties between different optimal assignments are broken arbitrarily.