

Adaptive experiments for policy choice (preliminary draft)

Maximilian Kasy* Anja Sautmann†

February 2, 2019

Abstract

The goal of many field experiments is to inform future policy choices. However, standard experimental designs are instead geared toward testing hypotheses. We consider the problem of optimal treatment assignment in an experiment with several cross-sectional waves. Our goal is to choose among a set of possible treatments for large-scale implementation. We show that optimal experimental designs learn from early waves by assigning more experimental units to the better-performing policy options 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, and also relative to standard Thompson sampling. Our setting is related to but different from multi-armed bandit settings; it does not involve an “exploitation” motive.

KEYWORDS: EXPERIMENTAL DESIGN, FIELD EXPERIMENTS, OPTIMAL POLICY

JEL CODES: C93, C11, O22

1 Introduction

Field experiments have become part of the standard toolkit of economics (Banerjee et al., 2016). Field experiments can serve various ultimate goals. Two possible goals

*Department of Economics, Harvard University, maximiliankasy@fas.harvard.edu

†Massachusetts Institute of Technology, sautmann@mit.edu.

We thank the participants at the development economics retreat at Harvard in winter 2018. All errors and opinions are our own.

are the estimation of treatment effects, and testing for the presence of a non-zero treatment effect. Standard recommendations for the design of experiments are based on these goals. These recommendations (see for instance Athey and Imbens 2017) include (i) assigning an equal number of units to different treatments, (ii) if possible within strata defined by pre-determined covariates, and (iii) choosing the sample size based on power calculations for tests of the hypothesis of a zero average treatment effect.

Another important goal is to inform future policy choices. We will show that the objective of informing policy leads to design recommendations that are qualitatively different from standard recommendations. This is in particular the case when (i) there are multiple alternative policies (treatments) that we wish to compare and (ii) the experiment can be run in multiple waves.

Example To fix ideas, consider the example of an NGO that has the goal of improving newborn health by encouraging “kangaroo care” (skin-to-skin care) for prematurely born babies. The NGO knows that kangaroo care, implemented correctly and used consistently, works in principle (Conde-Agudelo et al., 2012). There is, however, a considerable number of implementation choices to be made, which determine the uptake and quality of kangaroo care that is de-facto received, such as setting the right incentives for health-care providers, or educating mothers. How should the NGO make these choices? We argue that the NGO should run an experiment in multiple waves. Initially they should try many different variants. As they learn which options appear to perform better, they should focus their experiment on the best performing options. This allows them to learn the optimal option with high certainty. Once the experiment is concluded, they can then recommend the best performing option for large scale implementation. As this example suggests, one way to think about our proposal is that it gives 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, for the same sample size, relative to standard recommendations.

Setup More formally, we consider an experimental setting with multiple waves of experimental units, and multiple treatments. At the beginning of each wave, the number of units assigned to each treatment arm is decided. Then the assignments are implemented for this wave, and outcomes are observed, before treatments are assigned

for the next wave. Once the experiment is concluded, one of the treatments is picked for large-scale implementation. The goal is to maximize the average outcomes for this large-scale implementation, net of the costs of treatment. We would like to design experiments that are optimal or close to optimal for this goal.

Note that, while having some commonalities, this setting is different from the well-known “multi-armed bandit” problem (cf. Bubeck and Cesa-Bianchi, 2012; Weber et al., 1992). In particular, there is no “exploitation” motive, and thus no exploitation-exploration tradeoff in our setting.

Optimal assignment The setup just sketched defines a dynamic stochastic optimization problem. The actions in each wave (time period) are treatment assignments, the state space are beliefs over treatment effects, and transitions are determined by experimental outcomes. The last action to take is the choice of policy for large-scale implementation, after which welfare is realized as average outcomes net of costs. This dynamic stochastic optimization problem can, in principle, be solved analytically using backward induction. For some small-scale examples, we will discuss analytic solutions and their qualitative features. In more realistic settings, however, finding exact solutions quickly becomes computationally infeasible, due to exploding state and action spaces.

Modified Thompson sampling These computational constraints motivate the use of approximate solutions that are computationally feasible. We will in particular propose the following assignment algorithm, which is a modified version of so-called Thompson sampling (Russo et al., 2018). For standard 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 non-random 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. We will show, using simulations and theoretical results, that this modification improves expected welfare. It avoids assigning shares of more than 50% to the best-performing treatment, and in large samples it equalizes power

for rejecting each of the sub-optimal treatments.

Our baseline model does not take into account covariates. After discussing this baseline model, we consider covariates for stratified assignment algorithms and targeted treatment assignment policies. We propose, in particular, to use a hierarchical Bayes approach in order to form posterior beliefs about the treatment effectiveness for each treatment arm and stratum. This approach allows to learn about effect heterogeneity and optimal targeting, while at the same time importing information across strata in an optimal way.

Simulation evidence We provide extensive simulation evidence on the performance of alternative assignment algorithms. We evaluate these algorithms in terms of the welfare they generate, via informing a policy choice. Our simulations use parameters and sample sizes calibrated to data from published experiments in development economics (Ashraf et al., 2010; Bryan et al., 2014; Cohen et al., 2015).

Several patterns are confirmed by these simulations. Modified Thompson sampling consistently performs better than standard Thompson sampling, which in turn outperforms conventional non-adaptive designs. That is, the distribution of welfare (across simulations) under modified Thompson sampling stochastically dominates the alternatives. In particular, modified Thompson sampling generates higher average welfare, and a higher probability of picking the optimal treatment. The gains from adaptive 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.

Implementation in practice We next discuss a field experiment for which we implemented our recommended approach. [THIS IS WORK IN PROGRESS.]

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

The idea of adaptive experimental design has produced a large theoretical and practical literature, focused in particular on the so-called multi armed bandit problem. We will 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, rather than solving for the optimal assignment, heuristic algorithms are used far more commonly. Two popular algorithms are the Upper Confidence Bound algorithm (UCB), and Thompson sampling, cf. Russo et al. (2018). A large theoretical literature characterizes the expected regret of these algorithms, cf. Bubeck and Cesa-Bianchi (2012). Generalizations of the Bandit problem are discussed under the name of reinforcement learning in the machine learning literature, cf. 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.

Roadmap The rest of this paper is structured as follows. In Section 2, 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. In Section 4, we provide simulation evidence on the relative performance of adaptive algorithms, using parameters calibrated to data from published field experiments. In Section 5, we provide a theoretical analysis of the behavior of the proposed modified Thompson algorithm. In Section 6, we discuss inference for adaptive designs. In Section 7 we present the results of a field experiment where we implemented our approach, demonstrating its practical feasibility. [TBD] Section 8 concludes. The supplementary appendix also 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, in other words, 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 observations. 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

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

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

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

The experimenter's problem is to choose \mathbf{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

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

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

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

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 \mathbf{m}_t and \mathbf{r}_t .

Policy choice and welfare. After wave T , a policy $d^* \in 1, \dots, k$ 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

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

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

$$d^* = \operatorname{argmax}_d SW(d).$$

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 θ^d : $Y^d \sim \text{Ber}(\theta^d)$. The experiment and treatment assignment are designed to learn as quickly as possible about the individual success rates θ^d and choose the treatment with the highest θ^d .

We assume that the policymaker holds prior belief

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

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

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

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

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

Beta-Binomial distribution,

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

2.2 Optimal assignment

The choice of treatment assignment \mathbf{n}_t for each $t = 1, \dots, T$ 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 $(\mathbf{m}_{t-1}, \mathbf{r}_{t-1})$, and the action in t is given by \mathbf{n}_t . The transition between states is described by $\mathbf{m}_t = \mathbf{m}_{t-1} + \mathbf{n}_t$, $\mathbf{r}_t = \mathbf{r}_{t-1} + \mathbf{s}_t$, 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 $(\mathbf{m}_t, \mathbf{r}_t)$. After the experiment is concluded, the value function is given by the optimal choice of policy, based on current beliefs:

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

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

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

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

$$\begin{aligned} V_{t-1}(\mathbf{m}_{t-1}, \mathbf{r}_{t-1}) &= \max_{\mathbf{n}_t: \sum_d n_t^d \leq N_t} U_t(\mathbf{m}_{t-1}, \mathbf{r}_{t-1}, \mathbf{n}_t) \\ \mathbf{n}_t^*(\mathbf{m}_{t-1}, \mathbf{r}_{t-1}) &= \operatorname{argmax}_{\mathbf{n}_t: \sum_d n_t^d \leq N_t} U_t(\mathbf{m}_{t-1}, \mathbf{r}_{t-1}, \mathbf{n}_t). \end{aligned} \quad (2.4)$$

¹In 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 $\mathbf{c} = 0$ 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.

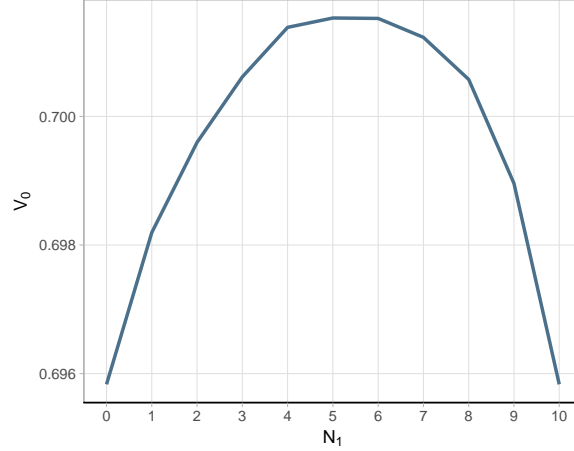
Figure 2.1 plots expected welfare as a function of the sample size N_1 in wave 1. The boundary cases $N_1 = 0$ and $N_1 = 10$ correspond to an experiment with only one wave. The optimal split assigns either 5 or 6 units to the first wave. Splitting the sample in this manner allows to learn from the first-wave 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 $(1 + s_1^d)/4$.

²The 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.

³The figures in Appendix A.2 explore similar scenarios for different sizes of wave N_2 .

Figure 2.1: Dividing the sample across waves.

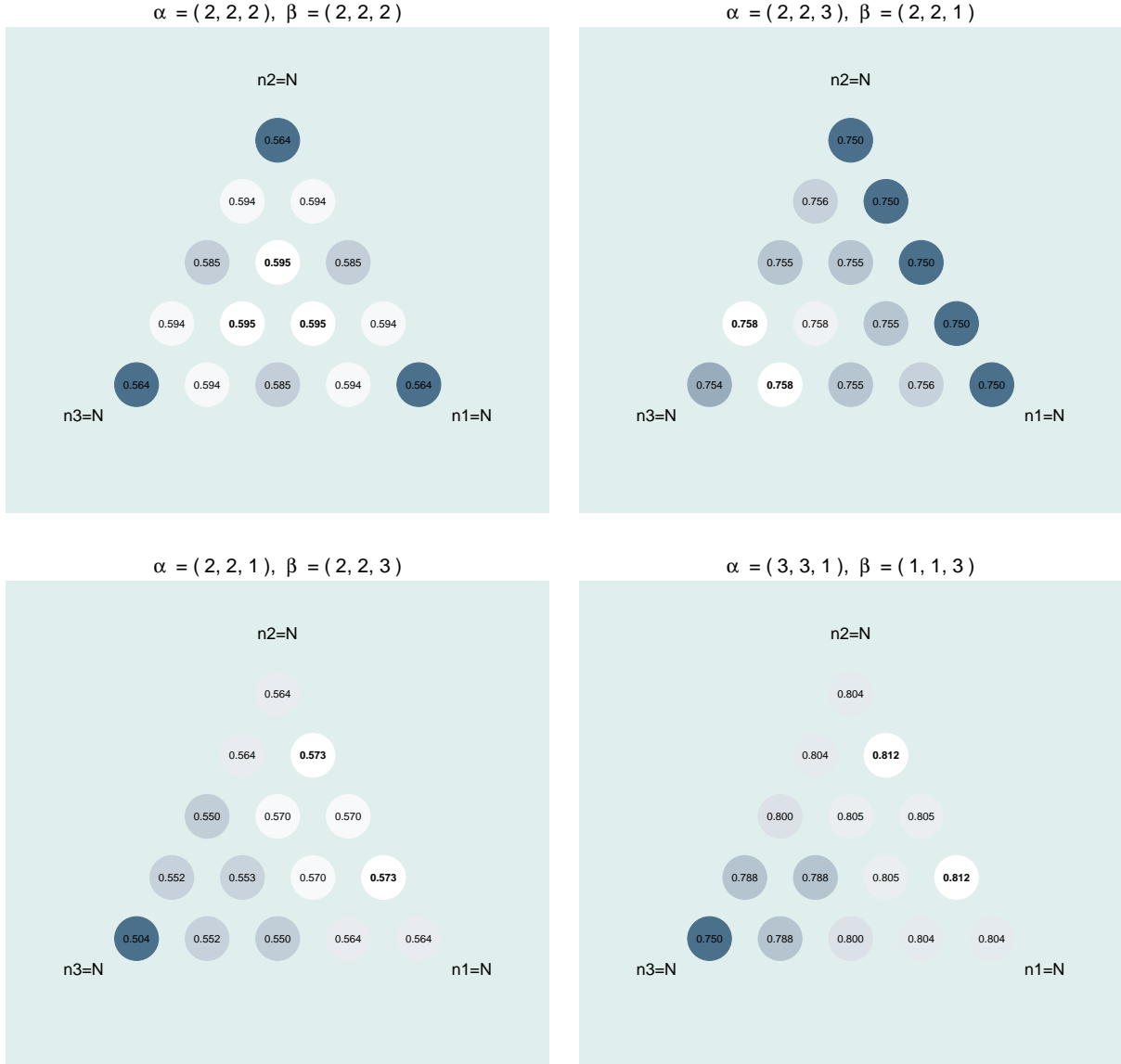


Notes: The graph shows expected welfare V_0 as a function of the sample size N_1 in period 1, assuming a total sample size of 10 and three treatments, for a uniform prior.

Four scenarios The four scenarios we consider are $\mathbf{s}_1 = (1, 1, 1)$, $\mathbf{s}_1 = (1, 1, 2)$, $\mathbf{s}_1 = (1, 1, 0)$, and $\mathbf{s}_1 = (2, 2, 0)$. 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 $\mathbf{n}_2 = (n_2^1 + n_2^2 + n_2^3)$ in wave 2 (which is of size 4), taking as given the Beta-prior parameters α_1, β_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 $d^* = \operatorname{argmax}_d E[\theta^d \mathbf{m}_T, \mathbf{r}_T]$
Loss function	$\sum_d (\hat{\theta}^d - \theta^d)^2$	$\max_d \theta^d - \theta^{d^*}$
Risk function	$\sum_d \operatorname{Var}_{\theta}(\hat{\theta}^d) + \operatorname{Bias}_{\theta}^2(\hat{\theta}^d)$	$\max_d \theta^d - E_{\theta}[\theta^{d^*}]$
Bayes risk	$\sum_d E[(\hat{\theta}^d - \theta^d)^2]$	$\max_d \theta^d - E[\max_d E[\theta^d \mathbf{m}_T, \mathbf{r}_T]]$

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 $\max_d (E[\theta^d | \mathbf{m}_T, \mathbf{s}_T] - c^d)$. 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 $\mathbf{c} = 0$.

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 $\text{Regret} = \max_d \theta^d - \theta^{d^*}$. 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

In the previous section we discussed fully optimal treatment assignments, and calculated them for some simple examples. In realistic settings, optimal solutions are computationally infeasible, however.

An alternative to full optimization is the use of simpler algorithms that approximate the optimal solution. Such approximate algorithms are widely used in the context of 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 conceived by Thompson (1933) in the context of clinical trials, Thompson sampling can be defined as follows.

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,

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

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

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

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, \dots, 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, \dots, 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*. 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.

3.1 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 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 consideration translates into a prior with statistical dependence between the vectors of average potential outcomes $\boldsymbol{\theta}^x = (\theta^{1x}, \dots, \theta^{kx})$ 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 θ^{dx} be the corresponding average potential outcome. Assume

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

where π is some prior distribution for (α^d, β^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 (α^d, β^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 θ^{dx} across strata, as reflected in (α^d, β^d) . Then use these as a prior, in conjunction with s_t^{dx} for a given stratum x , to learn about θ^{dx} for that stratum.

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

1. Gibbs step:

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

2. Metropolis steps:

- Given $\beta_{\rho-1}$ and θ_ρ , draw α_ρ^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 $\beta_{\rho-1}$ given θ_ρ and $\alpha_{\rho-1}$.

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 α , β and θ given $\mathbf{m}_t, \mathbf{r}_t$. In particular, we have that the posterior probability of a treatment d being optimal given x , in the sense of maximizing the probability of employment, is given by

$$P\left(d = \operatorname{argmax}_{d'} \theta^{d'x} | \mathbf{m}_t, \mathbf{r}_t\right) = \operatorname{plim}_{R \rightarrow \infty} \frac{1}{R} \sum_{\rho=1}^R \mathbf{1}\left(d = \operatorname{argmax}_{d'} \theta_\rho^{d'x}\right),$$

where θ^{dx} is the probability for a participant of type x to find employment when receiving treatment d .

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

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

For the empirical Bayes approach, we do not require a prior distribution over (α_0^d, β_0^d) . Instead, we generate “posterior” draws as follows:

1. First, for each d estimate (α^d, β^d) using **maximum likelihood**, based on the Beta-Binomial likelihood for the observed \mathbf{r}_{t-1}^d .
2. For each draw of (α^d, β^d) , sample a draw of θ^d from its Beta posterior given (α^d, β^d) and given $\mathbf{m}_{t-1}^d, \mathbf{r}_{t-1}^d$.

Then proceed as before.

4 Calibrated simulations

We next present simulation evidence on the performance of alternative treatment assignment algorithms, using parameter vectors θ 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.

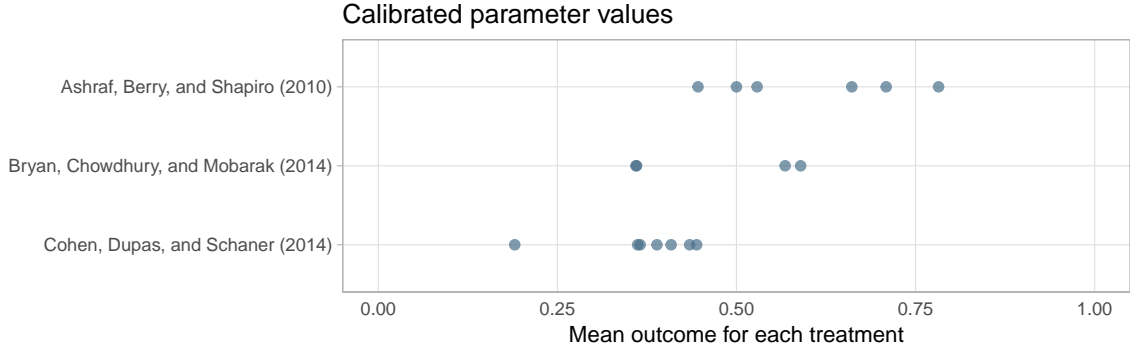
Bryan et al. (2014) conducted a field experiment in rural Bangladesh. Households were randomly assigned a cash or credit incentive of \$8.50, conditional on a household member migrating during the 2008 monga (lean) season. This amount covers the round-trip travel cost. The treatments in this experiment 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. $\mathbf{c} = 0$. This is of course unrealistic, but allows us to avoid ad-hoc assumptions regarding \mathbf{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, 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. 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 $\mathbf{c} = 0$ (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,

$$d^* = \operatorname{argmax}_d E[\theta^d | \mathbf{m}_T, \mathbf{r}_T],$$

$$\text{Regret} = \max_d \theta^d - \theta^{d^*}.$$

For each of our simulations the vector $\boldsymbol{\theta}$ is fixed, 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 θ^{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 $\text{Regret} = 0$.

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.1 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 $\text{Regret} = 0$) for each of the algorithms. 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 by the distribution using non-adaptive assignment. The integrated difference between the two quantile functions

is 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.

- The first pattern is that 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.
- The second pattern is that adaptive designs with more waves consistently outperform designs with fewer waves (for the same total sample size).
- The third pattern is that the gains from adaptive design 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 somewhat easier to detect.

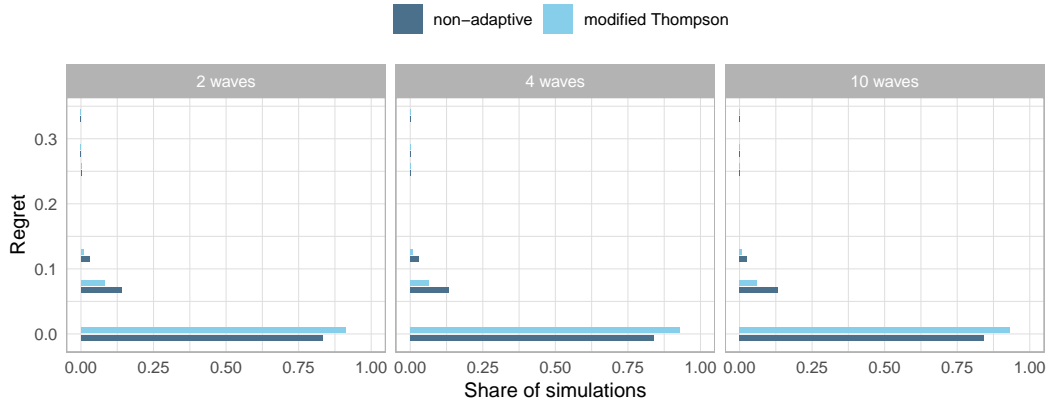
The figures reveal the following additional properties of modified Thompson sampling.

- The probability of choosing the best treatment is strictly larger than under non-adaptive assignment, for every setting considered.
- More generally, the distribution of regret under modified Thompson sampling first-order stochastically dominates the corresponding distribution under non-adaptive assignment.
- 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.

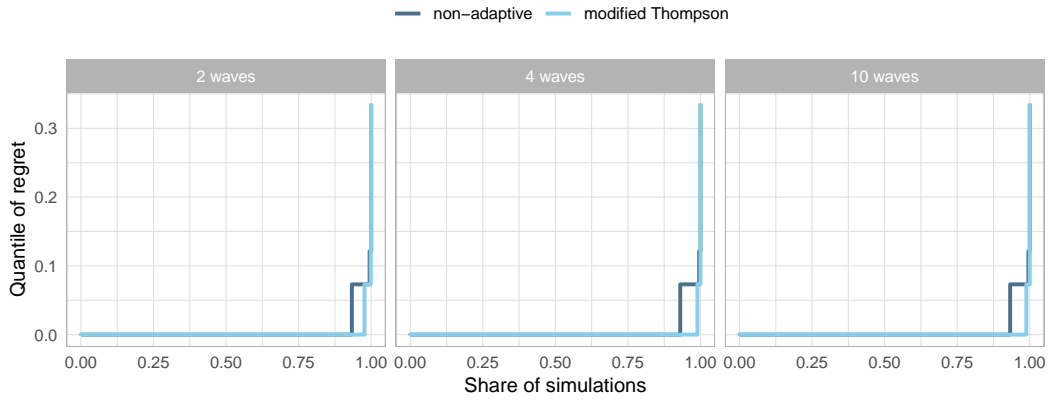
Table 2: Ashraf, Berry, and Shapiro (2010)

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

Ashraf, Berry, and Shapiro (2010)



Ashraf, Berry, and Shapiro (2010)

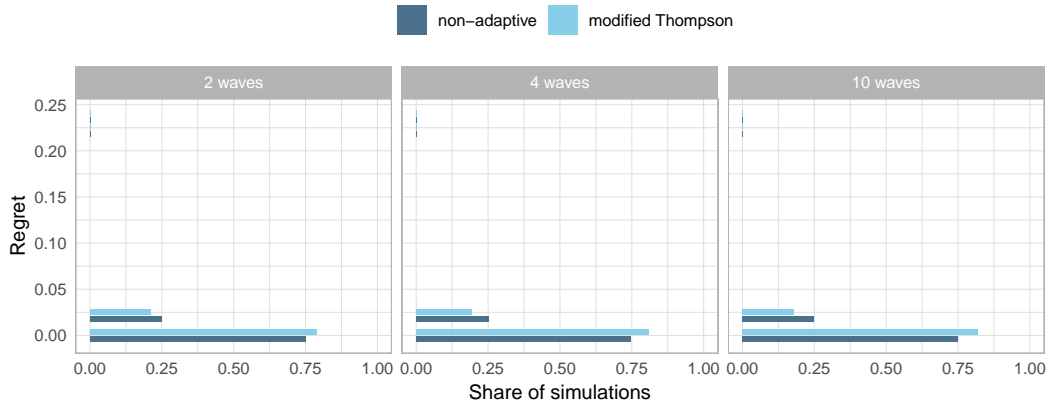


Notes: The table shows average regret and the share of replications for which the optimal treatment was chosen across 20,000 simulation replications. The histograms and quantiles show the distribution of regret. Parameters are calibrated based on the data of Ashraf et al. (2010). Total sample size is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

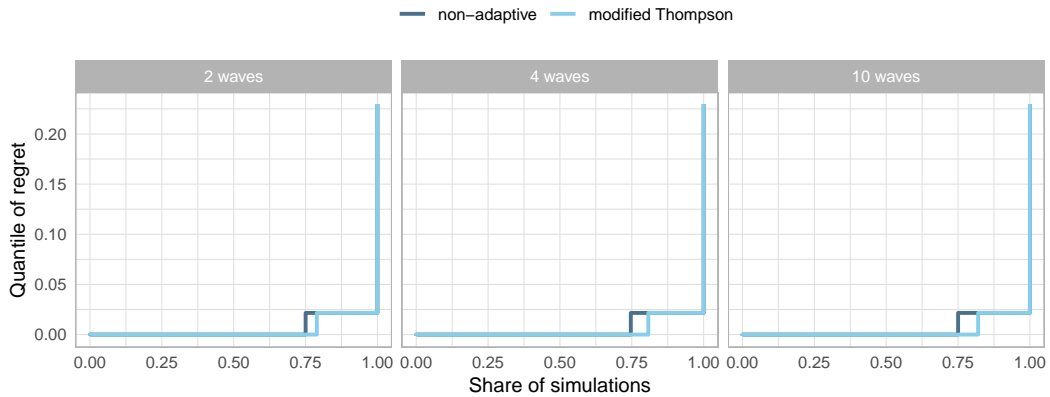
Table 3: Bryan, Chowdhury, and Mobarak (2014)

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

Bryan, Chowdhury, and Mobarak (2014)



Bryan, Chowdhury, and Mobarak (2014)

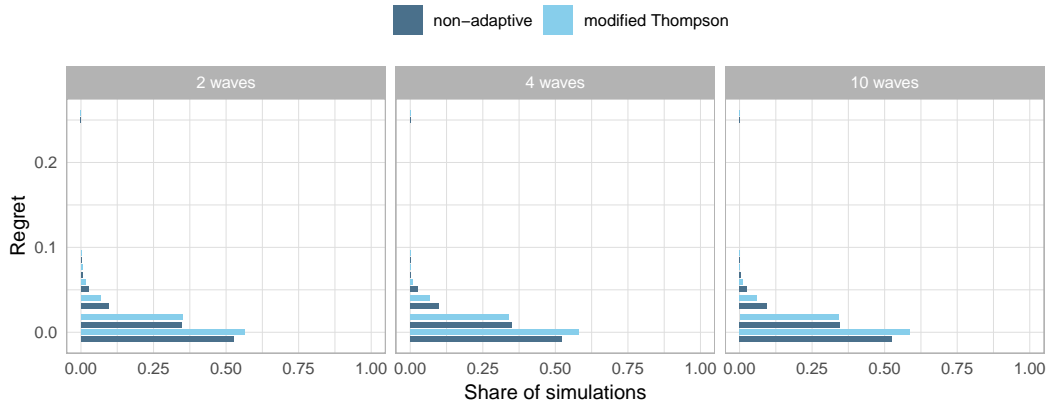


Notes: The table shows average regret and the share of replications for which the optimal treatment was chosen across 20,000 simulation replications. The histograms and quantiles show the distribution of regret. Parameters are calibrated based on the data of Bryan et al. (2014). Total sample size is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

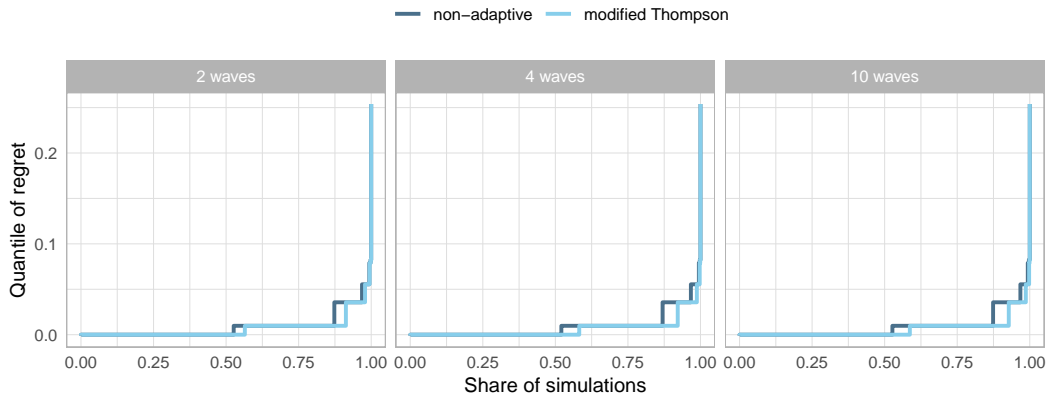
Table 4: Cohen, Dupas, and Schaner (2014)

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

Cohen, Dupas, and Schaner (2014)



Cohen, Dupas, and Schaner (2014)



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 used both for experimental treatment assignment, and for targeting of d^* , corresponding to the setting discussed in Section 3.1. We calibrate both the distribution across strata p_x and the conditional average potential outcomes θ^{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 θ^{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 θ 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 treatment assignment algorithms. The first algorithm is non-adaptive assignment, neglecting covariates, as before. 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 3.1. We use a default prior for the hyper-parameters (α_0^d, β_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, where we use 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

$$d^*(x) = \operatorname{argmax}_d E[\theta^{dx} | \mathbf{m}_T, \mathbf{r}_T] - c^d,$$

$$\text{Regret} = \sum_x p_x \left(\max_d \theta^{dx} - \theta^{d^*(x)} \right),$$

where p_x is the probability of covariate value x in the population, and $\mathbf{m}_T, \mathbf{r}_T$ 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 $\text{Regret} = 0$.

Simulation results The following tables show our simulation results for these settings, based on 5000 simulation draws.⁴ These tables are based on total sample sizes equal to the original experiments. The tables in Appendix A.1 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.

- Again, modified Thompson sampling performs better than Thompson sampling, which in turn dominates non-adaptive stratified assignment.
- Regret is larger than in the corresponding settings without covariates. This is owed 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.
- Correspondingly, the share of simulations for which the optimal policy was chosen is smaller.
- 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	0.012	0.010
expected Thompson	0.013	0.012	0.011
Thompson	0.013	0.012	0.012
non-adaptive stratified	0.017	0.018	0.017
non-adaptive	0.026	0.020	0.018
Share optimal			
modified Thompson	0.166	0.192	0.212
expected Thompson	0.166	0.199	0.198
Thompson	0.164	0.184	0.193
non-adaptive stratified	0.140	0.132	0.124
non-adaptive	0.086	0.125	0.137
Units per wave	502	251	100

⁴Given the larger computational cost for MCMC-based simulations in the hierarchical setting, we have a smaller number of simulation draws. This results in somewhat noisier estimates relative to the unconditional case.

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

Statistic	2 waves	4 waves	10 waves
Regret			
modified Thompson	0.005	0.005	0.004
expected Thompson	0.005	0.005	0.005
Thompson	0.005	0.004	0.005
non-adaptive stratified	0.007	0.007	0.007
non-adaptive	0.009	0.007	0.006
Share optimal			
modified Thompson	0.738	0.760	0.782
expected Thompson	0.702	0.735	0.754
Thompson	0.739	0.757	0.740
non-adaptive stratified	0.675	0.675	0.665
non-adaptive	0.602	0.651	0.685
Units per wave	935	467	187

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

Statistic	2 waves	4 waves	10 waves
Regret			
modified Thompson	0.016	0.015	0.015
expected Thompson	0.017	0.016	0.015
Thompson	0.017	0.016	0.015
non-adaptive stratified	0.019	0.019	0.019
non-adaptive	0.026	0.022	0.021
Share optimal			
modified Thompson	0.077	0.091	0.083
expected Thompson	0.065	0.082	0.082
Thompson	0.067	0.070	0.085
non-adaptive stratified	0.062	0.062	0.062
non-adaptive	0.029	0.048	0.050
Units per wave	1080	540	216

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 will 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 Before going into the details, let us provide an overview of the following arguments. We start off by heuristically 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 $\log(n)/n$, including Thompson sampling. What that implies, in particular, is that after n observations, only about $\log(n)$ 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 be able to distinguish the best option from sub-optimal options after the experiment. Thompson sampling stops learning about the performance of sub-optimal options too quickly, when the objective is policy choice.

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 options 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 probability, 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 we end up selecting the optimal policy with higher probability than under Thompson sampling, which in turn allows to select the optimal policy with higher probability than non-adaptive assignment, as reflected in the simulations of Section 4.

5.1 Heuristic motivation for modified Thompson sampling: Alternating treatment

Consider the algorithm described in the preview above, where we assign treatments based on the Thompson probabilities \mathbf{p} , but don’t assign the same treatment twice in a row. This algorithm defines a Markov chain. 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 \mathbf{q} , where the stationary distribution satisfies the equations

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

for $d \in \{1, \dots, k\}$. Moreover, the assignment shares of the “alternating” algorithm converge to the stationary distribution characterized by Equation (5.1). We can solve explicitly for \mathbf{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

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

5.2 The large sample behavior of Thompson sampling

5.3 The large sample behavior of modified Thompson sampling

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
θ^2	1	1	1	3/4
$E[\cdot \theta]$	θ	θ	$1/2 \cdot \theta(1 + \theta)$	$1/3 + \theta/6 + \theta^2/4$
$E_{\theta \sim U[0,1]}[\cdot]$	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, best-half assignment, Thompson sampling, and modified Thompson sampling.

We will show, using a simple example, that adaptive treatment assignment biases conventional estimators, such as sample means, and corresponding tests. We then show that Bayesian inference, however, 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.

Inference versus welfare maximization 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 θ^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. Suppose that we observe a binary random variable $Y_1 \sim \text{Ber}(\theta)$. If Y_1 is 0, we stop the experiment, if $Y_1 = 1$ we continue, and obtain another (independent) draw $Y_2 \sim \text{Ber}(\theta)$. 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 θ ! 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 ignoring the fact that sampling was adaptive remains valid. To illustrate, assume that we start with a uniform (i.e., $Beta(1, 1)$) prior for θ . Then the posterior mean $\hat{\theta}^B$ for θ 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 depended 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 θ . This contrasts with the prior mean of \bar{Y} , which is equal to $5/12$.

General validity of standard Bayesian inference This 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 $i = 1, 2, \dots, N$ with potential outcomes (Y_i^1, \dots, Y_i^k) . Denote the likelihood of Y_i^d , indexed by the parameter vector θ , by $f_{\theta}^d(Y_i^d)$.

Let $H_i = (D_1, \dots, D_{i-1}, Y_1, \dots, Y_{i-1})$ be the history of treatments and outcomes before i . Suppose that the treatment assigned to unit i is a function of H_i , as well as possibly a randomization device U_i .⁵

Proposition 1 Under Assumption 1, the likelihood of $(D_1, \dots, D_N, Y_1, \dots, Y_N)$ equals

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

up to a constant that does not depend on θ .

Proof: This follows immediately from the factorization

$$P(D_1, \dots, D_N, Y_1, \dots, Y_N | \theta) = \prod_{i=1}^N P(D_i | H_i, \theta) \cdot P(Y_i | D_i, H_i, \theta),$$

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

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.

⁵This setting includes as a special case treatment assignment in waves, and binary outcomes, as in Section 2.1.

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 $i = 1 \dots n$, drawing \tilde{U}_i from the same distribution as U_i , defining

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

and setting

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

The following proposition is immediate.

Proposition 2 *Consider the setting of Assumption 1.*

1. *Under the null hypothesis 1, \tilde{H}_n has the same distribution as H_n , conditional on (Y_1, \dots, Y_n) .*
2. *Let $T(H_n)$ be any test statistic, and let T^α be the $(1 - \alpha)$ quantile of the distribution of $T(\tilde{H}_n)$ over draws of $\tilde{U}_1, \dots, \tilde{U}_n$. Then a test which rejects iff $T(H_n) > T^\alpha$ controls size at level α .*

7 Implementation in the field

8 Conclusion

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 Trends® 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	Additional simulation results	36
A.1.1	No covariates	36
A.1.2	Targeting based on covariates	43
A.2	Additional optimal design example plots	46

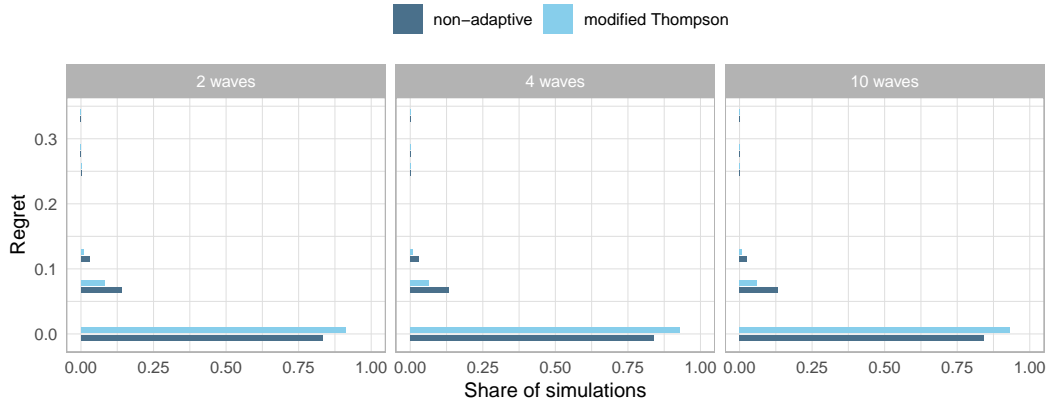
A.1 Additional simulation results

A.1.1 No covariates

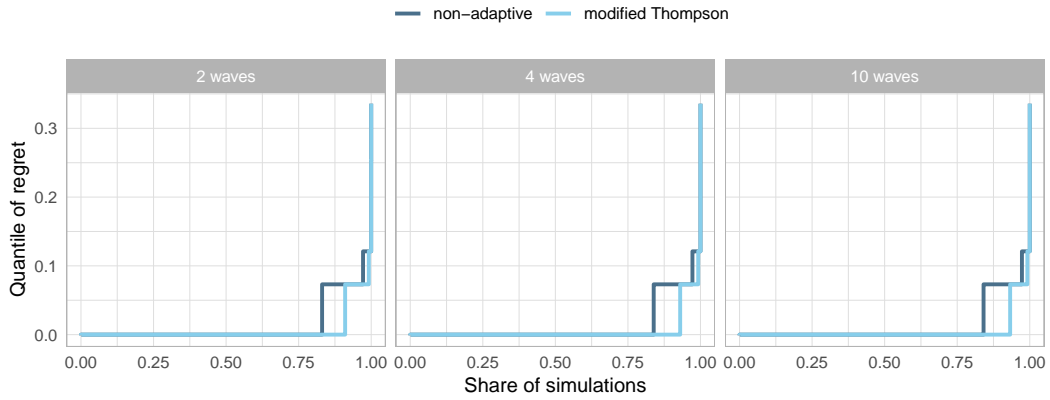
Table A1: Ashraf, Berry, and Shapiro (2010)

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

Ashraf, Berry, and Shapiro (2010)



Ashraf, Berry, and Shapiro (2010)

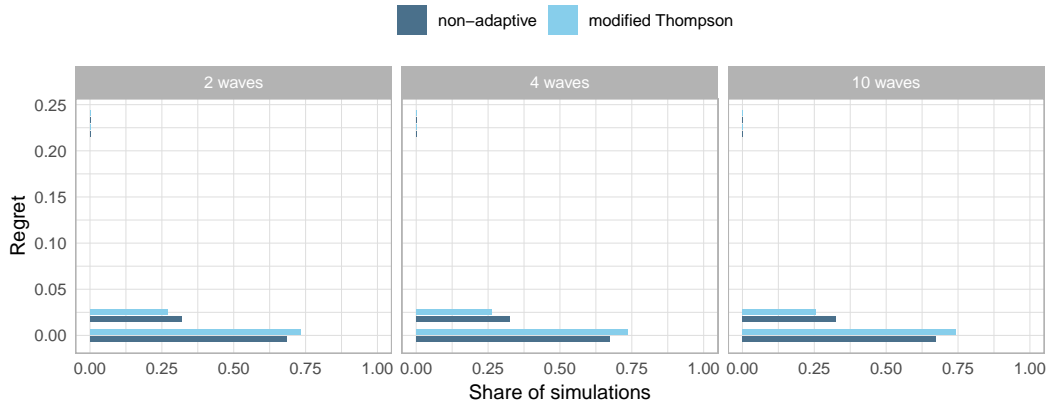


Notes: The table shows average regret and the share of replications for which the optimal treatment was chosen across 20,000 simulation replications. The histograms and quantiles show the distribution of regret. Parameters are calibrated based on the data of Ashraf et al. (2010). Total sample size is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

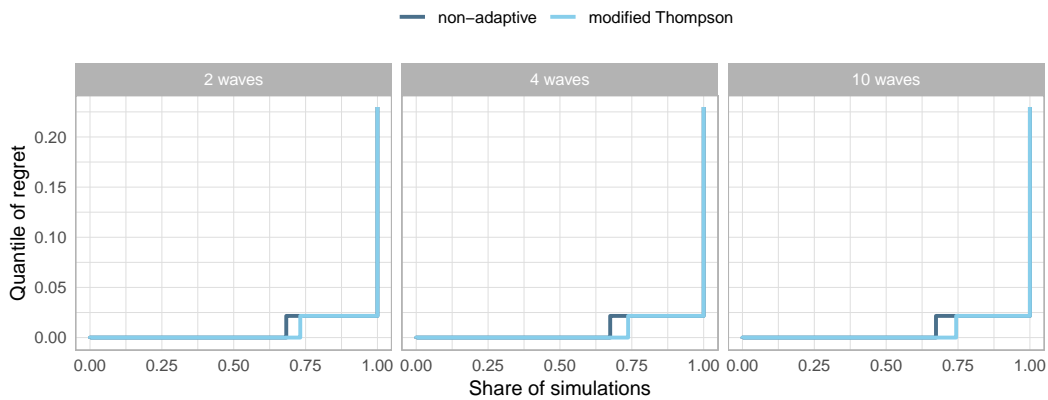
Table A2: Bryan, Chowdhury, and Mobarak (2014)

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

Bryan, Chowdhury, and Mobarak (2014)



Bryan, Chowdhury, and Mobarak (2014)

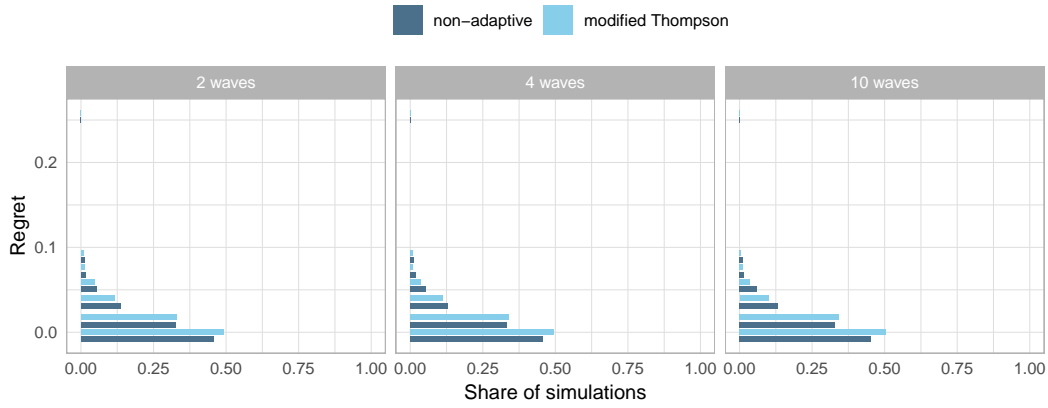


Notes: The table shows average regret and the share of replications for which the optimal treatment was chosen across 20,000 simulation replications. The histograms and quantiles show the distribution of regret. Parameters are calibrated based on the data of Bryan et al. (2014). Total sample size is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

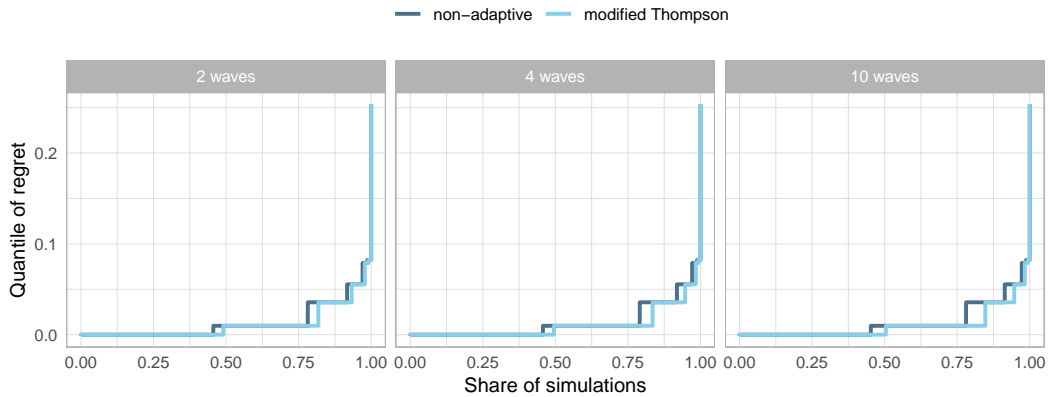
Table A3: Cohen, Dupas, and Schaner (2014)

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

Cohen, Dupas, and Schaner (2014)



Cohen, Dupas, and Schaner (2014)

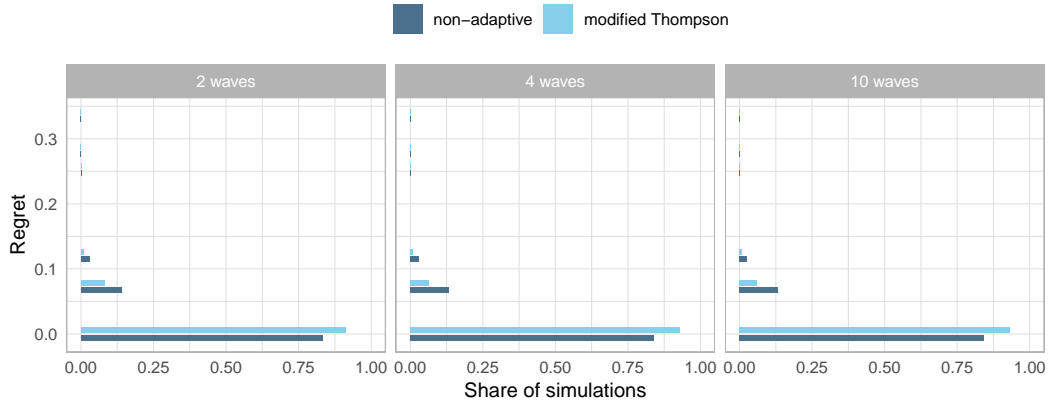


Notes: The table shows average regret and the share of replications for which the optimal treatment was chosen across 20,000 simulation replications. The histograms and quantiles show the distribution of regret. Parameters are calibrated based on the data of Cohen et al. (2015). Total sample size is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

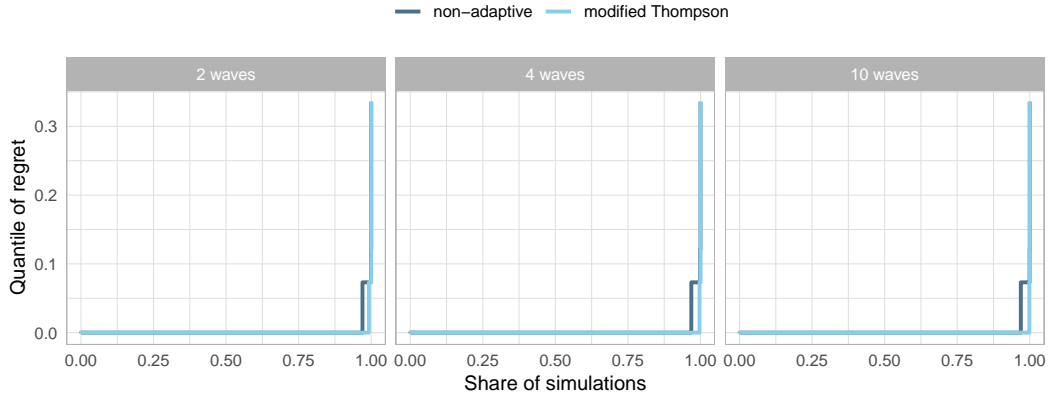
Table A4: Ashraf, Berry, and Shapiro (2010)

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

Ashraf, Berry, and Shapiro (2010)



Ashraf, Berry, and Shapiro (2010)

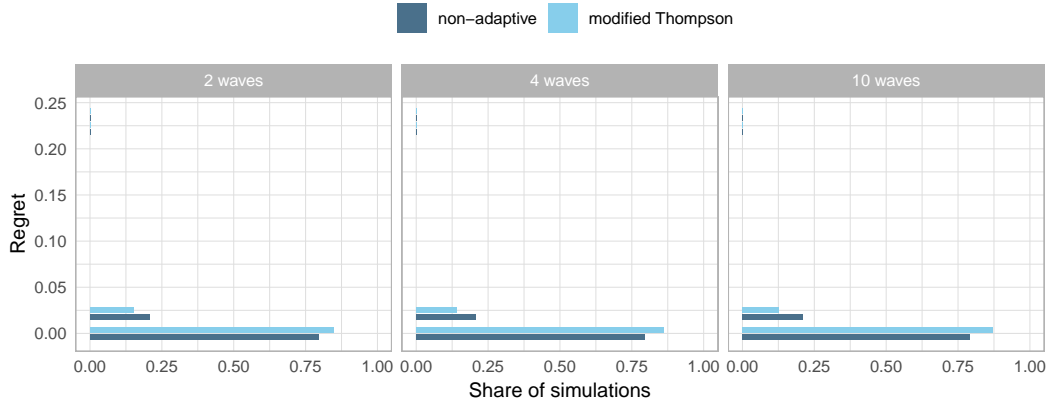


Notes: The table shows average regret and the share of replications for which the optimal treatment was chosen across 20,000 simulation replications. The histograms and quantiles show the distribution of regret. Parameters are calibrated based on the data of Ashraf et al. (2010). Total sample size is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

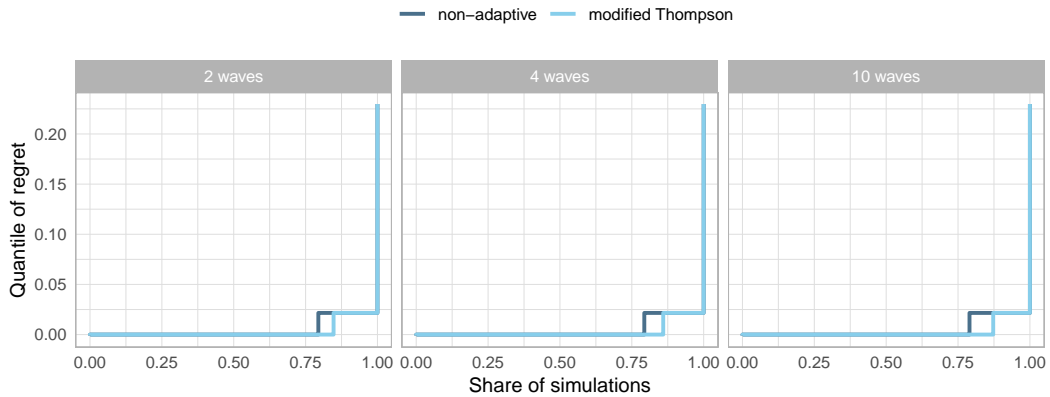
Table A5: Bryan, Chowdhury, and Mobarak (2014)

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

Bryan, Chowdhury, and Mobarak (2014)



Bryan, Chowdhury, and Mobarak (2014)

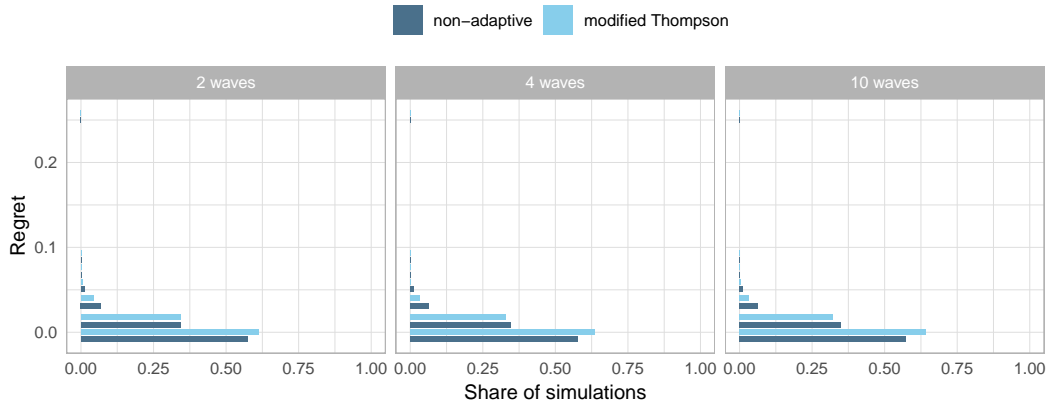


Notes: The table shows average regret and the share of replications for which the optimal treatment was chosen across 20,000 simulation replications. The histograms and quantiles show the distribution of regret. Parameters are calibrated based on the data of Bryan et al. (2014). Total sample size is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

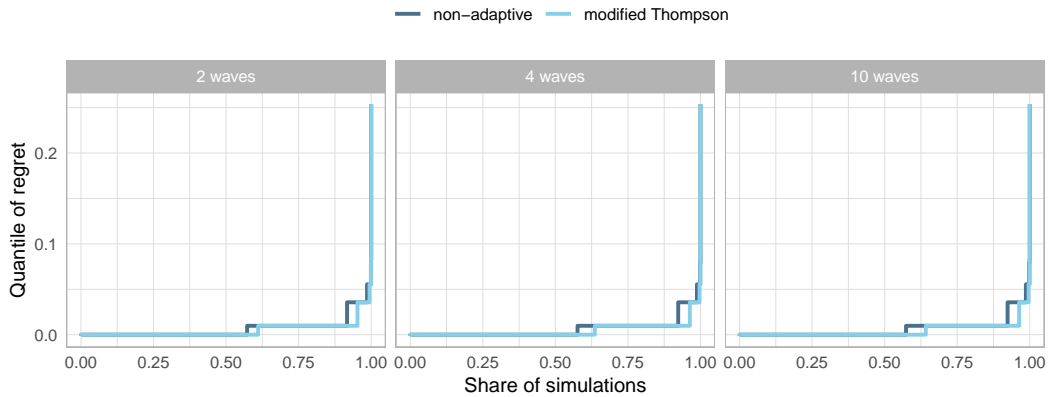
Table A6: Cohen, Dupas, and Schaner (2014)

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

Cohen, Dupas, and Schaner (2014)



Cohen, Dupas, and Schaner (2014)



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.1.2 Targeting based on covariates

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

Statistic	2 waves	4 waves	10 waves
Regret			
modified Thompson	0.024	0.021	0.020
expected Thompson	0.024	0.021	0.020
Thompson	0.023	0.020	0.020
non-adaptive stratified	0.028	0.029	0.029
non-adaptive	0.043	0.032	0.030
Share optimal			
modified Thompson	0.099	0.103	0.106
expected Thompson	0.102	0.110	0.121
Thompson	0.091	0.115	0.116
non-adaptive stratified	0.072	0.081	0.077
non-adaptive	0.044	0.066	0.073
Units per wave	251	125	50

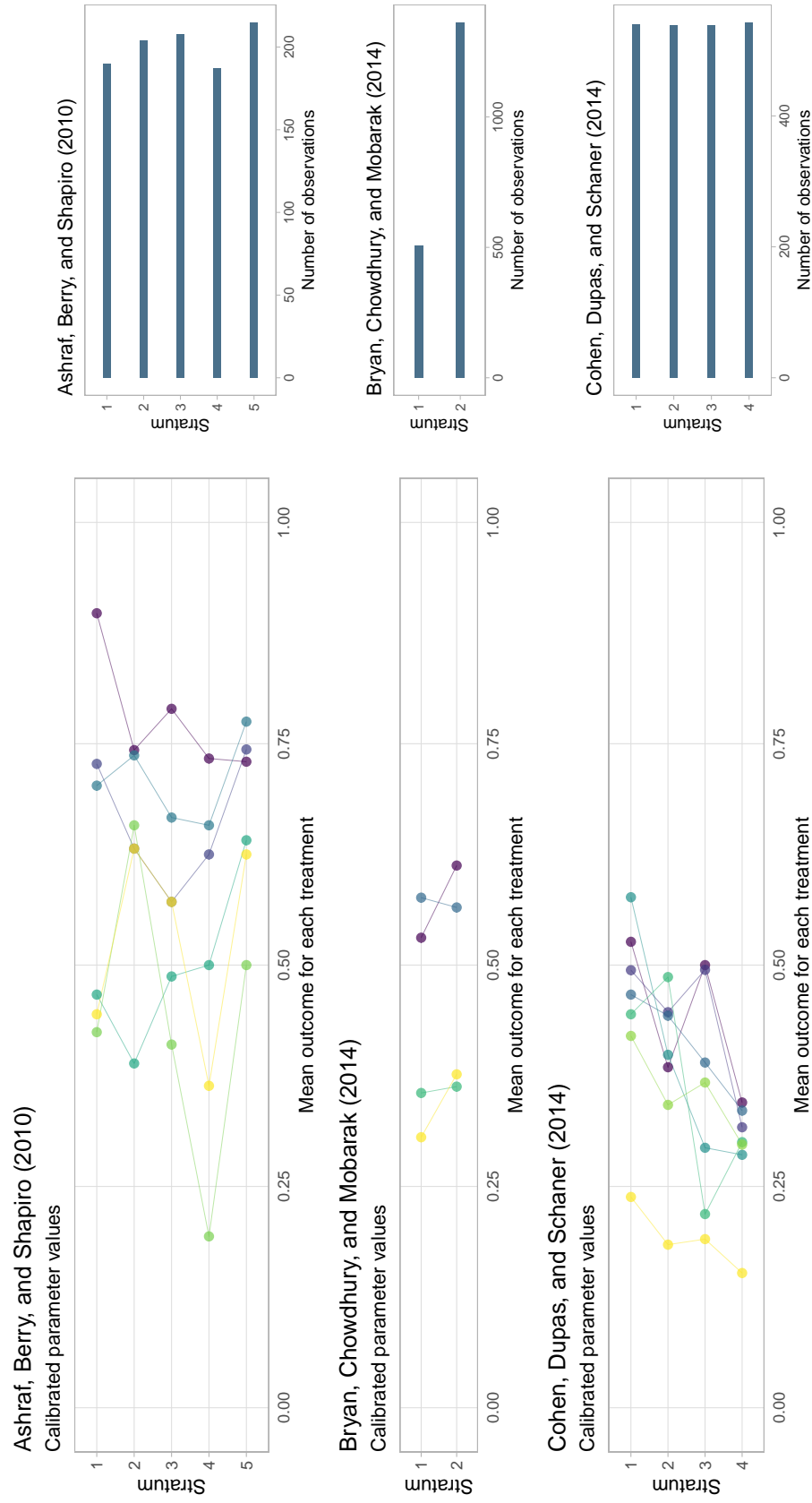
Table A8: Bryan, Chowdhury, and Mobarak (2014)

Statistic	2 waves	4 waves	10 waves
Regret			
modified Thompson	0.008	0.008	0.008
expected Thompson	0.010	0.008	0.008
Thompson	0.009	0.008	0.009
non-adaptive stratified	0.011	0.011	0.011
non-adaptive	0.013	0.012	0.011
Share optimal			
modified Thompson	0.606	0.601	0.636
expected Thompson	0.570	0.604	0.604
Thompson	0.600	0.612	0.583
non-adaptive stratified	0.528	0.529	0.535
non-adaptive	0.476	0.511	0.544
Units per wave	467	233	93

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

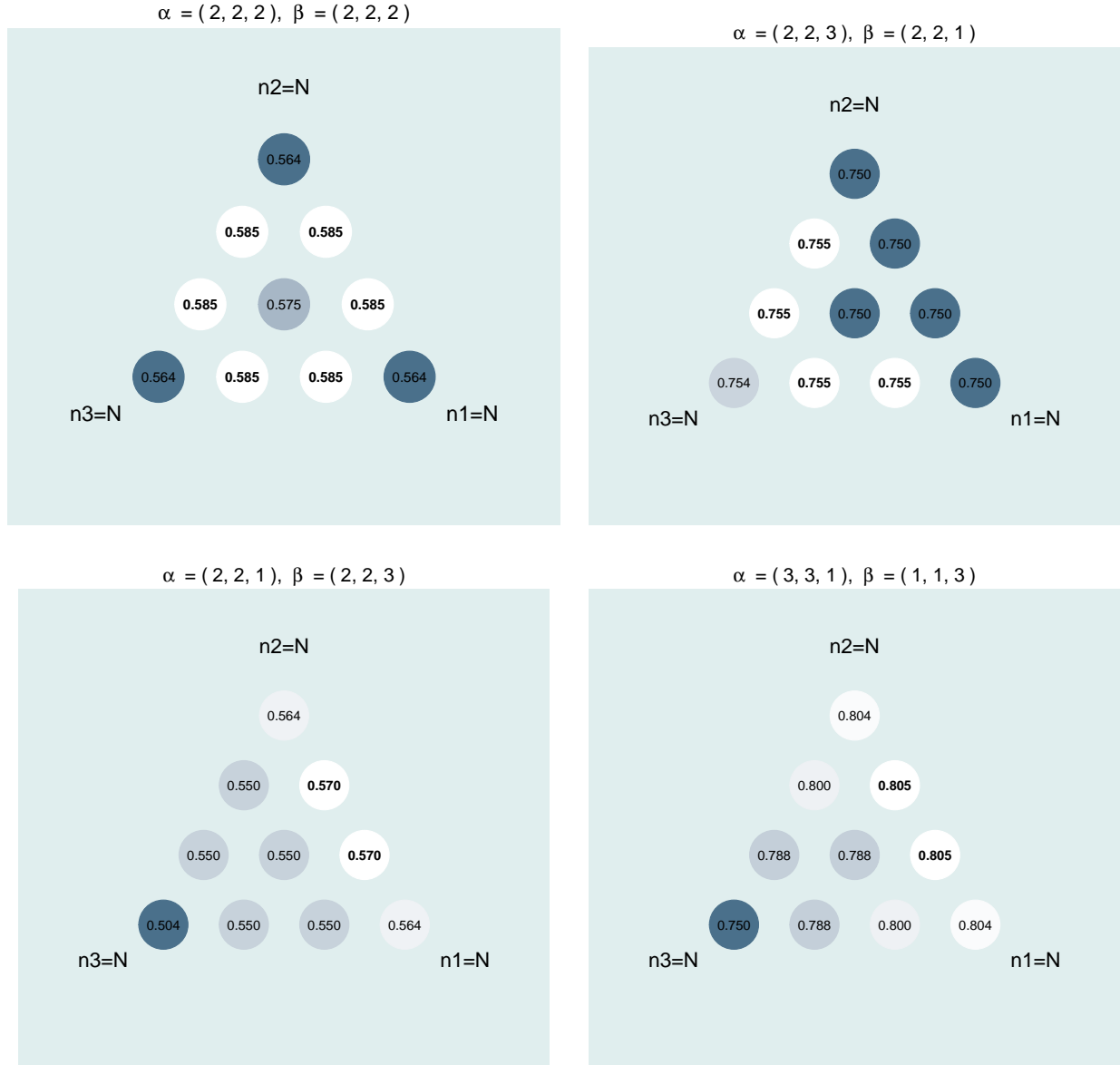
Statistic	2 waves	4 waves	10 waves
Regret			
modified Thompson	0.025	0.023	0.023
expected Thompson	0.025	0.023	0.023
Thompson	0.026	0.024	0.023
non-adaptive stratified	0.028	0.028	0.028
non-adaptive	0.035	0.030	0.028
Share optimal			
modified Thompson	0.037	0.045	0.046
expected Thompson	0.034	0.040	0.042
Thompson	0.034	0.037	0.044
non-adaptive stratified	0.027	0.027	0.025
non-adaptive	0.021	0.022	0.027
Units per wave	540	270	108

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



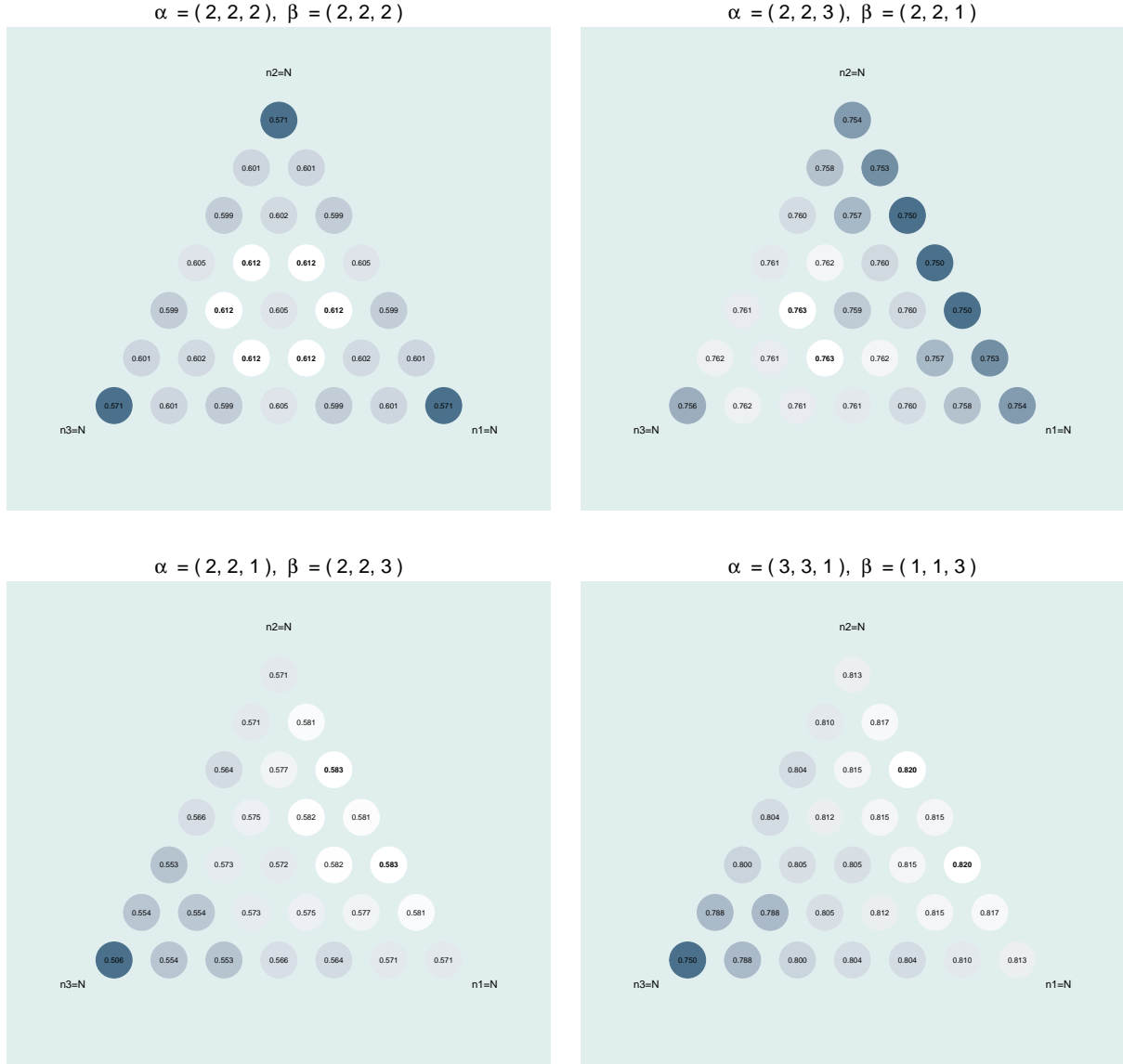
A.2 Additional optimal design example plots

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



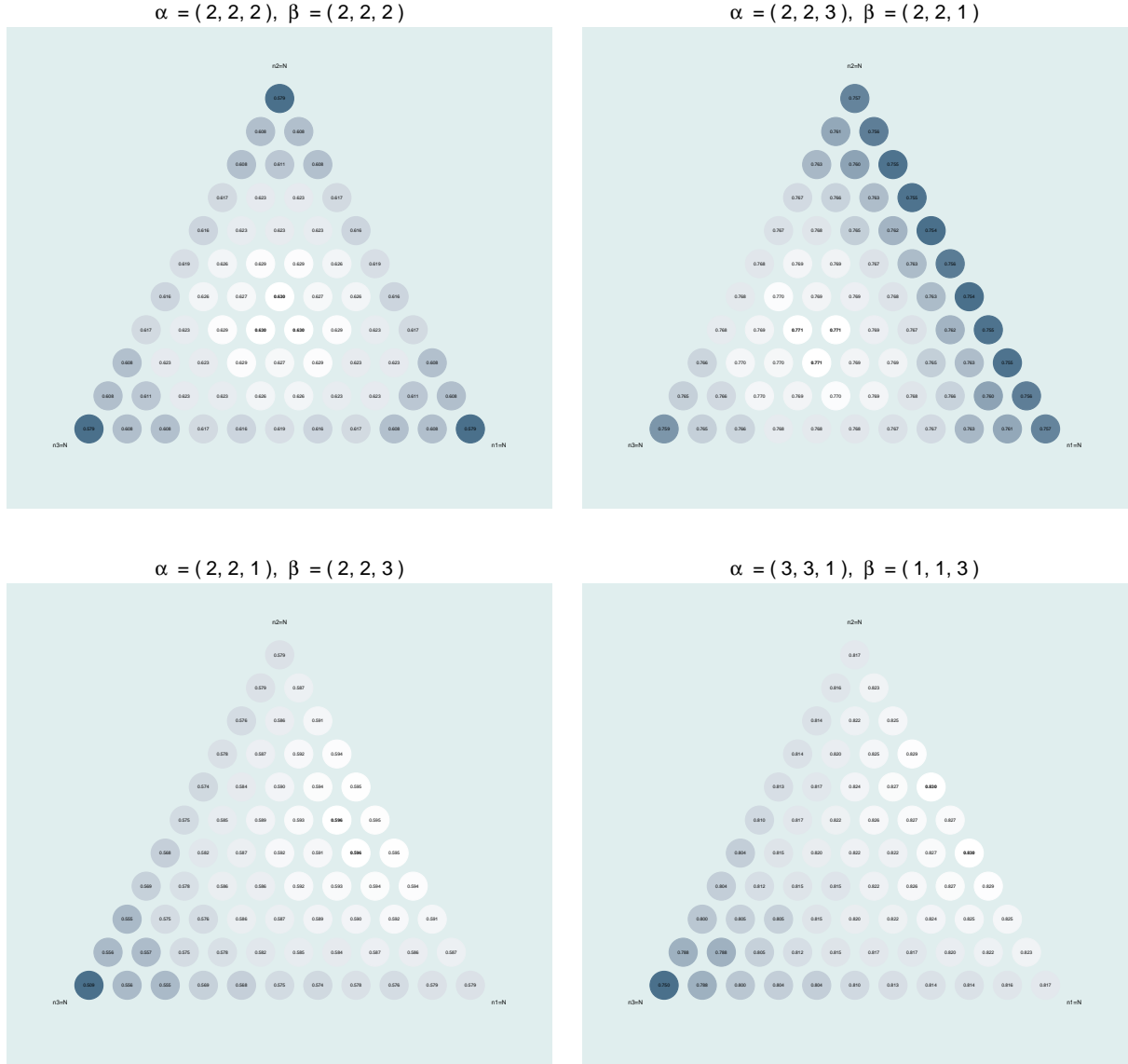
Notes: This figure shows the expected welfare U_2 as a function of treatment assignment n_2 in wave 2 (size 3), taking as given the *Beta*-prior parameters α_1, β_1 determined by the outcomes of wave 1 (size 6). Note that the color scaling differs across figures for better readability.

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 n_2 in wave 2 (size 6), taking as given the *Beta*-prior parameters α_1, β_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 n_2 in wave 2 (size 10), taking as given the *Beta*-prior parameters α_1, β_1 determined by the outcomes of wave 1 (size 6). Note that the color scaling differs across figures for better readability.