

Customer Acquisition via Display Advertising Using Multi-Armed Bandit Experiments

Eric M. Schwartz

Stephen M. Ross School of Business
University of Michigan

Eric Bradlow

Marketing Department
University of Pennsylvania

Peter Fader

Marketing Department
University of Pennsylvania

Ross School of Business Working Paper
Working Paper No. 1217
December 2013

This work cannot be used without the author's permission.

This paper can be downloaded without charge from the
Social Sciences Research Network Electronic Paper Collection:
<http://ssrn.com/abstract=2368523>

**Customer Acquisition via Display Advertising
Using Multi-Armed Bandit Experiments**

Eric M. Schwartz, Eric T. Bradlow, and Peter S. Fader *

December 2013

* Eric M. Schwartz is an Assistant Professor of Marketing at the Stephen M. Ross School of Business at the University of Michigan. Eric T. Bradlow is the K. P. Chao Professor; Professor of Marketing, Statistics, and Education; Vice Dean and Director of Wharton Doctoral Programs; and Co-Director of the Wharton Customer Analytics Initiative at the University of Pennsylvania. Peter S. Fader is the Frances and Pei-Yua Chia Professor; Professor of Marketing; and Co-Director of the Wharton Customer Analytics Initiative at the University of Pennsylvania. This paper is based on the first author's PhD dissertation. All correspondence should be addressed to Eric M. Schwartz: ericmsch@umich.edu, 734-936-5042; 701 Tappan Street, R5468, Ann Arbor, MI 48109-1234.

Customer Acquisition via Display Advertising Using Multi-Armed Bandit Experiments

Abstract

Online advertisers regularly deliver several versions of display ads in a single campaign across many websites in order to acquire customers, but they are uncertain about which ads are most effective. As the campaign progresses, they adapt to intermediate results and allocate more impressions to the better performing ads on each website. But how should they decide what percentage of impressions to allocate to each ad?

This paper answers that question, resolving the classic “explore/exploit” tradeoff using multi-armed bandit (MAB) methods. However, this marketing problem contains challenges, such as hierarchical structure (ads within a website), attributes of actions (creative elements of an ad), and batched decisions (millions of impressions at a time), that are not fully accommodated by existing MAB methods. We address this marketing problem by utilizing a hierarchical generalized linear model with unobserved heterogeneity combined with an algorithm known as Thompson Sampling. Our approach captures how the impact of observable ad attributes on ad effectiveness differs by website in unobserved ways, and our policy generates allocations of impressions that can be used in practice. We implemented this policy in a live field experiment delivering over 700 million ad impressions in an online display campaign with a large retail bank. Over the course of two months, our policy achieved an 8% improvement in the customer acquisition rate, relative to a control policy, without any additional costs to the bank. Beyond the actual experiment, we performed counter-factual simulations to evaluate a range of alternative model specifications and allocation rules in MAB policies.

Keywords: multi-armed bandit, online advertising, field experiments, adaptive experiments, sequential decision making, exploration/exploitation tradeoff, reinforcement learning, hierarchical models.

1 Introduction

Business experiments such as A/B or multivariate tests have become increasingly popular as a standard part of a firm’s analytics capabilities (Anderson and Simester 2011; Davenport 2009; Donahoe 2011; Urban et al. 2013). As a result, many interactive marketing firms are continuously *testing and learning* in their market environments. But as this practice becomes part of regular business operations, such sequential testing has to be done profitably: firms ought to be *earning while learning*.

One domain frequently using such testing is online advertising. For example, allocating all future impressions to the current “champion” ad – which, after testing multiple ad executions, has performed the best to date – could be a myopic strategy, also known as a *greedy* policy. By doing so, the firm may capitalize on chance, so it is not necessarily optimizing profits through learning. To reduce that risk, the firm may choose to test for a longer period, further exploring ad effectiveness before exploiting that information. We later refer to such a practice as a *test-then-learn* policy. Firms with limited advertising budgets commonly face this resource-allocation problem, so how should they decide what percentage of impressions to allocate to each ad on a week-by-week basis? In practice, an online advertiser commonly pre-purchases media, e.g., impressions, to be placed on a number of different websites, but it still needs to determine which ad to use for each of those impressions.

We focus on solving this problem, but first we emphasize that it is not unique to online advertisers; it belongs to a much broader class of sequential allocation problems that marketers have faced for years across countless domains. Many other activities – sending emails or direct mail catalogs, providing customer service, designing websites – can be framed as sequential and adaptive experiments. All of these problems are structured around the following questions: Which targeted marketing action should we take, when should we take it, and with which

customers and in which contexts should we test such actions?

This class of problems can be framed as a multi-armed bandit (MAB) problem (Robbins 1952; Thompson 1933). The MAB problem (formally defined later) is a classic adaptive experimentation and dynamic optimization problem. Some challenges of the associated business problems have motivated the development of various MAB methods. However, existing methods do not fully address the richness of the online advertising problem or many of the other aforementioned marketing problems. That is, the methods for solving the basic MAB problem, and even some generalizations, fall short of addressing common managerial issues.

The purpose of this paper is to shrink that gap. We aim to make two contributions, one substantive and one methodological. Substantively, we improve the practice of adaptive testing in online display advertising; methodologically, we extend existing methods to address a more general MAB problem, representative of a broader class of interactive marketing problems. To do so, we implemented our proposed MAB policy in a large-scale, adaptive field experiment in collaboration with a large retail bank focused on direct marketing to acquire customers. The field experiment generated data over two months in 2012, including more than 700 million ad impressions delivered across 84 different websites, which featured 12 unique banner ads that were described by three different sizes and four creative concepts. For comparison, we randomly assigned each observation to be treated by either our proposed MAB policy or a control policy (i.e., a balanced experiment). Using the data collected, we ran counterfactual policy simulations to understand how other MAB methods would have performed in this setting.

From a substantive perspective, we solve the problem facing firms buying online display advertising designed to acquire customers. How can you maximize customer acquisition rates by testing many ads on many websites while learning which ad works best on each website? Key insights include quantifying the value of accounting for attributes describing different ads

and unobserved differences (across websites) in viewers' responsiveness to those ad attributes. We glean these insights by using each website (more specifically, each media placement) as the unit of analysis in a heterogeneous model of acquisition from ad impressions. We introduce hierarchical modeling with partial pooling into our MAB policy, so we leverage information across all websites to allocate impressions across ads within any single website. In addition, while most existing MAB methods learn about each "arm" (e.g., ad creative) independently, we take an attribute-based approach that allows for "cross-arm" (e.g., cross-ad) learning, which speeds up the process of earning while learning.

The setting of online display advertising is particularly interesting because it presents three challenges. (1) While advertisers often try dozens of different ads, those ads are typically interrelated, varying by such attributes as creative design, message, and size/format. Thus, observing one ad's performance can suggest how similar ads will perform. (2) The way those attributes affect the ad's performance depends upon the context, such as the website on which it appears. (3) As is common in media buying, the advertiser's key decision is what percentage of the next batch of already-purchased impressions should be allocated to each ad, which is referred to as *rotation weights* used by the publisher. Our MAB method overcomes these three challenges commonly faced by firms.

From a methodological perspective, we propose a method for a version of the MAB that is new to the literature: a hierarchical, attribute-based, and batched MAB policy. The most critical component is unobserved heterogeneity (i.e., hierarchical, partially-pooled model). While some recent work has incorporated attributes into actions and batched decisions (Chapelle and Li 2011; Dani et al. 2008; Rusmevichientong and Tsitsiklis 2010; Scott 2010), no prior work has considered a MAB with action attributes and unobserved heterogeneity, which are central to the practical problem facing an online advertiser. Our proposed policy improves performance

(relative to benchmark methods) because we include unobserved heterogeneity in ad effectiveness across websites via partial pooling. In addition, we illustrate how incorporating (rather than ignoring) that heterogeneity leads to substantively different recommended allocations of resources.

To solve that problem, we propose an approach based on the principle known as *Thompson Sampling* (Thompson 1933) (also known as *randomized probability matching*, Chapelle and Li 2011; Granmo 2010; May et al. 2011; Scott 2010), but we extend existing methods to account for unobserved heterogeneity in the attribute-based MAB problem with batched decision making. The principle of Thompson Sampling is simply stated: The current proportion of resources allocated to a particular action should equal the current probability that the action is optimal (Thompson 1933). Beyond discussing how the proposed approach is conceptually different from prior work (Agarwal et al. 2008; Bertsimas and Mersereau 2007; Hauser et al. 2009), we also show via numerical experiments that it empirically outperforms existing methods in this setting. From a theoretical perspective, Thompson Sampling may seem like a simple heuristic, but it has recently been shown to be an optimal policy with respect to minimizing finite-time regret (Agrawal and Goyal 2012; Kaufmann et al. 2012), minimizing relative entropy (Ortega and Braun 2013), and minimizing Bayes risk consistent with a decision-theoretic perspective (Russo and Roy 2013).

The rest of the paper is structured as follows. Section 2 surveys the landscape of the substantive problem, online display advertising and media buying, from industry and research perspectives, as well as some background on MAB methods. This provides motivation for the extended MAB framework we describe. In Section 3, we translate the advertiser’s problem into MAB language, formally defining the MAB. Then in Section 4, we describe our two-part approach to solving the full MAB problem: a hierarchical (heterogeneous) generalized linear

model and the Thompson Sampling allocation rule. To contrast this problem (and our method) with existing versions of the MAB, we describe how existing methods would only solve simpler versions of the advertising problem. In particular, the *Gittins index policy* is an optimal solution to the Markov decision process of what we refer to as the “basic MAB problem” (Gittins 1979; Gittins and Jones 1974). However, the Gittins index only solves a narrow set of problems with restrictive assumptions (e.g., all ads are assumed to be unrelated). Therefore, we discuss why it is an inappropriate MAB policy for the ad allocation problem of interest in this paper and why Thompson Sampling is a more appropriate method to use here.

In Section 5, we turn to the empirical context. We illustrate institutional and implementation details about our field experiment with a large retail bank and then discuss the actual results. In Section 6, we consider what might have happened if we had used other MAB methods in the field experiment. These counterfactual policy simulations reveal which aspects of the novel method account for the improved performance. In particular, we split our analysis into evaluations of the two key components: the model and the allocation rule. Finally, in Section 7, we conclude with a general discussion of issues at the intersection of MAB methods, online advertising, customer acquisition, and real-time, scalable optimization of business experiments.

2 Online Display Advertising and Bandit Problems

We build on two main areas of the literature: online advertising and multi-armed bandit problems. Despite the common concerns about the (presumed) ineffectiveness of display ads, advertisers still use them extensively for the purpose of acquiring new customers. In 2012, combining display, search, and video, U.S. digital advertising spending was \$37 billion (eMarketer 2012b). Out of that total, display advertising accounted for 40%. Display advertising’s share of U.S. digital advertising spending is growing, and is expected to be greater than spon-

sored search by 2016 (eMarketer 2012b). While display advertising may be used to build brand awareness, it is also purchased to generate a direct response (e.g., sign-up, transaction, or acquisition) from those impressions. Indeed, direct-response campaigns were the most common purpose of display advertising campaigns and accounted for 54% of display advertising budgets in the U.S. and 67% in the U.K. in 2011 (eMarketer 2012a). Research in this area has focused on the impact of exposure to ads on purchases (Manchanda et al. 2006) or other customer activities like search behavior (Reiley et al. 2011).

In contrast to previous work (Manchanda et al. 2006; Reiley et al. 2011), we focus on optimizing the allocation of resources over time across many ad creatives and websites by sequentially learning about ad performance. This challenge relates to other work at the intersection of online advertising or online content optimization and MAB problems (Agarwal et al. 2008; Scott 2010; Urban et al. 2013). Like those studies, we downplay what the firm is specifically learning about ad characteristics (e.g., Is one color better than others for ads? Are tall ad formats better than wide ones?) in favor of addressing, “How can we learn which ads are most effective as profitably as possible?” In our particular setting, as mentioned, we solve the problem of how to allocate previously purchased impressions (i.e., the firm bought a set number of impressions for each website). But the general approach can be extended to the problem of where to buy media and other sequential resource allocations common to interactive marketing problems.

The marketing literature has had limited exposure to MAB problems and almost no exposure to methods beyond the dynamic programming literature, such as the Gittins index solution (Hauser et al. 2009; Lin et al. 2013; Meyer and Shi 1995; Urban et al. 2013). We recognize that Gittins (1979) solved a classic sequential decision making problem that attracted a great deal of attention (Berry 1972; Bradt et al. 1956; Robbins 1952; Thompson 1933; Wahrenberger

et al. 1977) and was previously thought to be intractable (Berry and Fristedt 1985; Gittins et al. 2011; Tsitsiklis 1986; Whittle 1980).

While the Gittins index has been applied to some extent in marketing and management science (Bertsimas and Mersereau 2007; Hauser et al. 2009; Urban et al. 2013), these same applications note that the Gittins index only solves a special case of the MAB problem with restrictive assumptions.¹ We echo those sentiments to reiterate that the Gittins index is inappropriate for our ad allocation problem. While our problem does not have an exact solution using dynamic programming methods, we can still evaluate the performance of different policies, as we do empirically using a set of benchmark policies.

Nevertheless, the bandit literature is large and spans many fields, so our aim is not a complete review (Gittins et al. 2011; White 2012). But as we noted earlier, currently existing MAB methodologies do not adequately address all of the ad problem’s key components. Therefore, an alternative MAB method is needed to address those challenges and is described next.

3 Formalizing Online Display Advertising as a Multi-Armed Bandit Problem

We translate the advertiser’s problem into MAB language, formally defining the MAB and our approach to solving it with all components of the problem included. To contrast this problem with existing versions of the MAB, we will briefly describe how current MAB methods would only solve simpler versions of the advertising problem.

The MAB problem facing advertisers bears little resemblance to that basic bandit problem. In the problem addressed here, the firm has ads, $k = 1, \dots, K$, that it can serve on any or all of a set of websites, $j = 1, \dots, J$. Let impressions be denoted by m_{jkt} and conversions, by y_{jkt} ,

¹Hauser et al. (2009) applies the Gittins index to a webmorphing problem, adapting a website’s content based on a visitor’s inferred cognitive style. The MAB policy used assumes each morph (action) is independent, but it does account for unobserved heterogeneity via latent classes. Therefore, the Expected Gittins Index is a weighted average of the class-specific Gittins index over the class membership probabilities, which is an approximation shown in Krishnamurthy and Wahlberg (2009).

from ad k on website j in period t . Each ad's unknown conversion rate, μ_{jk} , is assumed to be stationary over time, but is specific to each website-ad combination.

The ad conversion rates are not only unknown, but they may be correlated since they are functions of unknown common parameters denoted by θ , and a common set of d ad attributes. Hence, the MAB is *attribute-based*. This attribute structure is X , which is the design matrix of size $K \times d$, where x_k corresponds to the k th row. To further emphasize the actions' dependence on those common parameters, we sometimes use the notation $\mu_{jk}(\theta)$, but we note that μ_{jk} is really a function of both x_k and a subset of parameters in θ .

For each decision period and website, the firm has a budget of $M_{jt} = \sum_{k=1}^K m_{jkt}$ impressions. In the problem we address, this budget constraint is taken as given and exogenous due to previously arranged media contracts, but the firm is free to decide what proportion of those impressions will be allocated to each ad. This proportion is w_{jkt} , where $\sum_{k=1}^K w_{jkt} = 1$. Since many observations are allocated simultaneously instead of one observation at-a-time, the problem is a *batched* MAB (Chick and Gans 2009).

In each decision period, the firm has the opportunity to make different allocations of impressions across K ads on each of J different websites. This ad-within-website structure implies the problem is *hierarchical*. Since each ad may perform differently depending on which website it appears, we allow an ad's conversion rate to vary by using website-specific attribute importance parameters, β_j . Then the impact of the ad attributes on the conversion rate can be described by a common generalized linear model (GLM), $\mu_{jk}(\theta) = h^{-1}(x_k' \beta_j)$, where h is the link function (e.g., logit, probit). Later, we specify the form of heterogeneity across websites. In particular, we will use a partial pooling approach, assuming all websites come from the same broader population. Intuitively, we can think of each website as a different slice of the population of all Internet traffic.

3.1 MAB Optimization Problem

The firm's objective is to maximize the expected total number of customers acquired by serving impressions. Like any dynamic optimization problem, the MAB problem requires the firm to select a *policy*. We define a MAB policy, π , to be a decision rule for sequentially setting allocations, \mathbf{w}_{t+1} , each period based on all that is known and observed through periods $1, \dots, t$, assuming f, h, K, X, J, T and \mathbf{M} are given and exogenous. That is, π maps information onto allocation of resources across actions. We aim to select a policy, π , that corresponds to an allocation schedule, \mathbf{w} , to maximize the cumulative sum of expected rewards, as follows,

$$\max_{\mathbf{w}} \mathbf{E}_f \left[\sum_{t=1}^T \sum_{j=1}^J \sum_{k=1}^K Y_{jkt} \right] \text{ subject to } \sum_{k=1}^K w_{jkt} = 1, \forall j, t, \quad (3.1)$$

where $\mathbf{E}_f [Y_{jkt}] = w_{jkt} M_{jt} \mu_{jk}(\theta)$.

Equation 3.1 lays out the undiscounted finite-time optimization problem, but we can also write the discounted infinite-time problem if we assume a geometric discount rate $0 < \gamma < 1$, let $T = \infty$, and maximize the expected value of the summations of $\gamma^t Y_{jkt}$. However, we will continue on with the undiscounted finite-time optimization problem, except where otherwise mentioned.

We can also express this optimization problem in a Bayesian decision-theoretic framework by specifying the likelihood of the data $p(Y|\theta)$ and prior $p(\theta)$,

$$\max_{\mathbf{w}} \int_{\theta} \sum_{t=1}^T \sum_{j=1}^J \sum_{k=1}^K Y_{jkt} \cdot p(Y_{jkt}|\theta) p(\theta) d\theta \text{ subject to } \sum_{k=1}^K w_{jkt} = 1, \forall j, t. \quad (3.2)$$

Due to the curse of dimensionality (Powell 2011), this problem does not correspond to an exact value function satisfying the Bellman optimality equations. However, there is marriage between dynamic programming and reinforcement learning seen in the correspondence

between the error associated with a value function approximation (i.e., Bellman error) and regret (i.e., opportunity cost expressed as $\Delta_k = \mu_{k^*} - \mu_k$, where k^* is the optimal arm) (Osband et al. 2013). While we do not provide a theoretical analysis deriving Thompson Sampling as a solution method to MAB problems, we refer the reader to work supporting its use (Kaufmann et al. 2012; Ortega and Braun 2010; Russo and Roy 2013).

To emphasize why we need a MAB policy to be more flexible than traditional approaches used in management sciences, we illustrate the special case of the MAB problem that would be solved exactly by a Gittins index policy. We would have to dramatically simplify and make largely unrealistic assumptions about our current advertising allocation problem to reduce it to the basic MAB problem. For instance, we would have to suppose there are no ad attributes (X is the identity matrix) yielding K independent and uncorrelated actions, there is just one single website ($J = 1$), and each batch contains only one observation ($M_{jt} = 1$). In addition, we would have to apply geometric discounting to rewards over an infinite-time horizon ($0 < \gamma < 1$, $T = \infty$; Gittins 1979). However, in our setting, we have to accommodate the aforementioned components of the marketing problem, and next we propose the policy that we use to overcome those real-world challenges.

4 MAB Policies

4.1 Thompson Sampling with a Hierarchical Generalized Linear Model

Now that we have described the advertising allocation problem as a hierarchical, attribute-based, batched MAB problem, we can focus on our MAB approach. While it is not an exact solution in the traditional, dynamic programming sense, it is a policy supported by a growing body of empirical evidence and theoretical guarantees of Thompson Sampling’s superior performance (Agrawal and Goyal 2012; Chapelle and Li 2011; Granmo 2010; Kaufmann et al.

2012; May et al. 2011; Ortega and Braun 2010; Russo and Roy 2013; Scott 2010).

Our proposed bandit policy is a combination of a *model*, the hierarchical generalized linear model (HGLM), and an *allocation rule*, Thompson Sampling. The specific HGLM of customer acquisition is a logistic regression model with varying parameters across websites. The Thompson Sampling allocation rule is straight-forward: the probability that an action is believed to be optimal is the proportion of resources that should be allocated to that action (Thompson 1933). This is achieved by drawing posterior samples from any model, which encodes parameter uncertainty. As a result, one draws actions randomly in proportion to the posterior probability that the action is the optimal one, encoding policy uncertainty. Before describing how we take advantage of Thompson Sampling in our setting, we describe the model of conversions accounting for display ad attributes and unobserved heterogeneity across websites.

Let the data collected in each time period, t , be fully described by the number of conversions, y_{jkt} , out of impressions, m_{jkt} , delivered per period, per ad for each of the J websites (contexts) and K ads (actions). We summarize our hierarchical logistic regression with covariates and varying slopes,

$$\begin{aligned}
 y_{jkt} &\sim \text{binomial}(\mu_{jk} | m_{jkt}) \\
 \mu_{jk} &= 1 / [1 + \exp(-x'_k \beta_j)] \\
 \beta_j &\sim \mathbf{N}(\bar{\beta}, \Sigma) \\
 x_k &= (x_{k1}, \dots, x_{kd}),
 \end{aligned} \tag{4.1}$$

where $\{\beta_j\}_1^J = \{\beta_1, \dots, \beta_J\}$, and all parameters are contained in $\theta = (\{\beta_j\}_1^J, \bar{\beta}, \Sigma)$. We could use a fully Bayesian approach with Markov Chain Monte Carlo simulation to obtain the joint posterior distribution of θ . However, for implementation in our large-scale real-time experi-

ment, we rely on the Laplace approximation to obtain posterior draws (Bates and Watts 1988), as is done in other Thompson Sampling applications (e.g., Chapelle and Li 2011). After obtaining estimates using restricted maximum likelihood, we perform model-based simulation by sampling parameters from the multivariate normal distribution implied by the mean estimates and estimated variance-covariance matrix of those estimates (Bates et al. 2013; Gelman and Hill 2007).

To simplify notation in subsequent descriptions of the models and allocation rules, we denote all conversions and impressions we have observed through time t as the set $\{\mathbf{y}_t, \mathbf{m}_t\} = \{y_{jk1}, m_{jk1}, \dots, y_{jkt}, m_{jkt} : j = 1, \dots, J; k = 1, \dots, K\}$. To include the attribute design matrix, we denote all data through t as, $D_t = \{X, \mathbf{y}_t, \mathbf{m}_t\}$. After an update at time t , we utilize the uncertainty around parameters β_j to obtain the key distribution for Thompson Sampling, the joint predictive distribution of conversion rates (expected rewards), $p(\mu_j | D_t)$.²

The preceding paragraphs described the HGLM, but this is one piece of the proposed MAB policy. The other piece is the allocation rule: Thompson Sampling (TS). Hence, we refer to the proposed MAB policy as TS-HGLM. In order to translate the predictive distribution of conversion rates into recommended allocation probabilities, $\mathbf{w}_{j,k,t+1}$, for each ad on each website in the next period, we apply the principle of TS, which works with the HGLM as follows. In short, for each website, j , we compute the probability that each of the K actions is optimal for that website and use those probabilities for allocating impressions. More formally, we obtain the distribution $p(\mu_j | D_t)$, and we can carry through our subscript j and then follow the procedures from the TS literature (Chapelle and Li 2011; Granmo 2010; May et al. 2011; Scott 2010). For each j , suppose the optimal action's mean is $\mu_{j*} = \max\{\mu_{j1}, \dots, \mu_{jK}\}$ (e.g.,

²Although we express the website-specific vector of the mean reward of each action as a K -dimensional vector, $\mu_j = (\mu_{j1}(\theta), \dots, \mu_{jK}(\theta))$, it is important to note that the reward distributions of the actions are not independent. The reward distributions of the actions in any website j are correlated through the common set of attributes X and common website-specific parameter β_j .

the highest true conversion rate for that website). Then we can define the set of allocation probabilities,

$$w_{j,k,t+1} = \Pr(\mu_{jk} = \mu_{j*} | D_t) = \int_{\mu_j} \mathbf{1}\{\mu_{jk} = \mu_{j*} | \mu_j\} p(\mu_j | D_t) d\mu_j, \quad (4.2)$$

where $\mathbf{1}\{\mu_{jk} = \mu_{j*} | \mu_j\}$ is the indicator function of which ad has the highest conversion rate for website j . The key to computing this probability is conditioning on μ_j and integrating over our beliefs about μ_j for all J websites, conditional on all information D_t through time t .

Since our policy is based on the HGLM, we depart from other applications of Thompson Sampling because our resulting allocations are based on a partially pooled model. While our notation shows separate \mathbf{w}_{jt} and μ_j for each j , those values are computed from the parameters β_j , which are partially pooled. Thus, we are not obtaining the distribution of β_j separately for each website; instead, we leverage data from all websites to obtain each website's parameters. As a result, websites with little data (or more within-website variability) are shrunk toward the population mean parameter vector, $\bar{\beta}$, representing average ad attribute importance across all websites. This is the case for all hierarchical models with unobserved continuous parameter heterogeneity (Gelman et al. 2004; Gelman and Hill 2007). Given those parameters, we use the observed attributes, X , to determine the conversion rates' predictive distribution, $p(\mu_j | D_t)$. For this particular model, the integral in Equation 4.2 can be rewritten as,

$$w_{j,k,t+1} = \int_{\Sigma} \int_{\bar{\beta}} \int_{\beta_1, \dots, \beta_J} \mathbf{1}\left\{\beta_j x_k = \max_k \beta_j x_k | \beta_j, X\right\} p(\beta_j | \bar{\beta}, \Sigma, X, \mathbf{y}_t, \mathbf{m}_t) p(\bar{\beta}, \Sigma | \beta_1, \dots, \beta_J) d\beta_1 \dots d\beta_J d\bar{\beta} d\Sigma. \quad (4.3)$$

However, it is much simpler to interpret the posterior probability, $\Pr(\mu_{jk}(\theta) = \mu_{j*}(\theta) | D_t)$, as a direct function of the joint distribution of the means, $\mu_j(\theta)$. It is natural to compute

allocation probabilities via posterior sampling (Scott 2010).³ We can simulate independent draws $g = 1, \dots, G$ of β_j . Each $\beta_j^{(g)}$ can be combined with the design matrix to form $\mu_j^{(g)} = h^{-1}(X' \beta_j^{(g)})$. Then, conditional on the g th draw of the K predicted conversion rates, the optimal action is to select the ad with the largest predicted conversion rates, $\mu_{j*}^{(g)} = \max\{\mu_{j1}^{(g)}, \dots, \mu_{jK}^{(g)}\}$. Across G draws, we approximate $w_{j,k,t+1}$ by computing the fraction of simulated draws in which each ad, k , is predicted to have the highest conversion rate,

$$w_{j,k,t+1} \approx \hat{w}_{j,k,t+1} = \frac{1}{G} \sum_{g=1}^G \mathbf{1} \left\{ \mu_{jk}^{(g)} = \mu_{j*}^{(g)} \mid \mu_j^{(g)} \right\}. \quad (4.4)$$

Computed from the data through periods $1, \dots, t$, the allocation weights, $\hat{w}_{j,k,t+1}$, combine with, $M_{j,t+1}$, the total number of pre-purchased impressions to be delivered on website j across all K ads in period $t + 1$. Since the common automated mechanism (e.g., DoubleClick for Advertisers) delivering the display ads does so in a random rotation according to the allocation weights, $(\hat{w}_{j,1,t+1}, \dots, \hat{w}_{j,K,t+1})$, the allocation of impressions is a multinomial random variable, $(m_{j,k,t+1}, \dots, m_{j,K,t+1})$, with the budget constraint $M_{j,t+1}$. However, since the number of impressions in the budget is generally very large in online advertising, each observed $m_{jkt} \approx M_{jt} \hat{w}_{jkt}$.

To conclude the description of our HGLM-TS policy, we provide additional justification for our use of TS. As mentioned earlier, while TS is a computationally efficient and simple allocation rule, it is not a suboptimal heuristic. Reinforcement learning and statistical learning work evaluates the performance of any MAB method by examining how the policy minimizes finite-time regret (error rate) asymptotically (Auer et al. 2002; Lai 1987). Recent research demonstrates that TS reduces that regret at an optimal rate (Agrawal and Goyal 2012; Kauf-

³In the case of two-armed Bernoulli bandit problem, there is a closed-form expression of the probability of one arm's mean is greater than the other's (Berry 1972; Thompson 1933).

mann et al. 2012). Other avenues of research focus on a fully Bayesian decision-theoretic perspective, and such approaches reach similar conclusions, finding TS minimizes Bayes risk (Russo and Roy 2013), as well as minimizing entropy when facing uncertainty about policy (Ortega and Braun 2010, 2013). After all, posterior sampling encodes the uncertainty in current beliefs, and the dynamics in the MAB stem exclusively from the explore/exploit tradeoff of how to learn profitably.

One major benefit of TS is that it is compatible with any model. Given a model’s predictive distribution of the arm’s expected rewards, it is straightforward to compute the probability of each arm having the highest expected reward. This means that we can examine a range of model specifications, just as we would ordinarily do when analyzing a dataset, and we can apply the TS allocation rule to each of those models. Later, we consider a series of benchmark MAB policies, including alternative models less complex than our HGLM and a set of alternative allocation rules instead of TS. For now, we turn to the empirical section and examine the performance of our proposed TS-HGLM policy in the field.

5 Field Experiment

5.1 Design and Implementation

We implemented a large-scale MAB field experiment by collaborating with the aforementioned bank and its online media-buying agency. The bank had already planned an experiment to test $K = 12$ new creative executions for their display ads using a two-factor design with three different ad sizes and four different ad concepts. The ad sizes were industry standards (160x600, 300x250, and 728x90 pixels), and the ads were delivered across $J = 84$ media placements. However, not every ad size was available on each website, but every ad concept is always available on each website regardless of ad sizes. We use the terms “website” and

“media placement” interchangeably for simplicity of exposition. We also use “acquisition” and “conversion” (i.e., from visitor to customer) interchangeably.

The goal of the test was to increase customer acquisition rates during the campaign. Previously, the bank had been running tests for a pre-set period of time (e.g., two months). After the test, they had been measuring ad performance using ad click-through rate in aggregate, across all websites. However, following our TS-HGLM policy, we changed this in four ways: (i) we measured performance using customer acquisition (not clicks); (ii) we looked at a more disaggregate level by analyzing ad performance website-by-website reflecting heterogeneity in β_j ; (iii) we changed the allocations each period, approximately every week (for $T = 10$ periods), following our proposed MAB policy based on ad performance throughout the campaign, and (iv) we simultaneously tested two policies (our TS-HGLM policy and a static balanced experiment).

To summarize the first three changes, we reiterate that while the campaign’s goal, maximizing customer acquisition, is simply stated, it involves learning which ad has the best acquisition rate for each media placement (e.g., website). The bank had already set its schedule over two months, as it purchased M_{jt} impressions for each media placement to be delivered each week, but the banked wanted to optimize its allocation of those impressions to ads within each media placement to maximize their return on their media purchases.

The fourth change reflects our desire as researchers to measure the impact of our proposed adaptive MAB policy compared to a control policy (e.g., static balanced experiment) in a real-time test. Therefore, the field experiment can be viewed as two parallel and identical hierarchical attribute-based batched MAB problems, with one treatment and one control group, where their only difference was the policy used to solve the same bandit problem. In the static group, we used an equal-allocation policy across ads (e.g., experiment with balanced design).

In the adaptive group, we ran the proposed algorithm, TS with a heterogeneous logit model (TS-HGLM). The same ads were served across all of the same websites over the same time period. The only difference was due to how our MAB policy allocated impressions between ads within any website for each time period. In the initial period, both groups (adaptive and static) were treated with the same equal-allocation policy. So any observed differences between the groups' acquisition rates in the initial period can be attributed to random binomial variation.

Throughout this empirical portion of this paper, all conversion rates reported are rescaled versions of the actual data from the bank (per the request of the firm to mask the exact customer acquisition data). We performed this scaling by a small factor, so it has no effect on the relative performance of the policies, and it is small enough so that almost all values of interest are within the same order of magnitude as their actual observed counterparts. In addition, we assign anonymous identities to media placements ($j = 101, \dots, 184$), ad sizes (A, B, and C), and ad concepts (1, 2, 3, and 4).

5.2 Field Experiment Results

To compare the two groups, we examine how the overall acquisition rate changed over time. The difference can be attributed to the proposed MAB (treatment) policy. On average, we expect the static (control) group's aggregate acquisition rate to remain flat. By contrast, we expect the rate for the adaptive (TS-HGLM) group to increase on average over time as the MAB policy "learns" which ad is the best ad k^* for each website j . Figure 1 provides evidence in support of those predictions, showing the results of the experiment involving more than 700 million impressions, where reallocations using TS-HGLM were made on each of the $J = 84$ websites every five to seven days over the span of two months, yielding a total of $T = 10$ periods. We examine the cumulative conversion rates at each period t , aggregated

across all ads and websites, computed as aggregate conversions $\sum_{\tau=1}^t \sum_{j=1}^J \sum_{k=1}^K y_{jk\tau}$ divided by aggregate impressions $\sum_{\tau=1}^t \sum_{j=1}^J \sum_{k=1}^K m_{jk\tau}$. We compare this cumulative conversion rate to the conversion rate during the initial period of equal allocation to show a percentage increase (as well as masking the firm’s true acquisition rates).

To summarize the key result, compared to the static balanced design, the TS-HGLM policy (the solid line in Figure 1) improves overall acquisition rate by 8%. In terms of customers, the economic impact of this treatment policy compared to the control can be stated as follows: We acquired approximately 240 additional new customers beyond the 3,000 new customers acquired through the control policy.

[INSERT FIGURE 1 ABOUT HERE]

From a substantive perspective, we note that these extra conversions come at no additional cost because the total media spend did not increase. They are the direct result of adaptively reallocating already-purchased impressions across ads within each website. Therefore, the cost per acquisition (CPA) decreases (CPA equals total media spend divided by number of customers acquired). In essence, we increased the denominator of this key performance metric by 8%. The new CPA has consequences beyond the gains during the experiments; it provides guidance for future budget decisions (e.g., how much the firm is willing to spend for each expected acquisition). We return to this in the general discussion when we show potential linkages to post-acquisition activities such as customer lifetime value.

5.3 A Closer Look at How the Thompson Sampling Works

We examine how the TS-HGLM policy works by focusing on the three aspects of the policy: (1) how allocations differ across combinations of ad concept and ad size (i.e., attributes), (2) how allocations differ across websites (i.e., heterogeneity), and (3) how allocations across ads

within a website change over time (i.e., learning). Consider two representative websites ($j = 103$ and 149) highlighted at two time points ($t = 1$ and 6) in Figure 2. The boxplots and lines summarize the predictive marginal distributions of the corresponding $\mu_{jk}(\theta)$ based on the HGLM.

[INSERT FIGURE 2 ABOUT HERE]

We see ad attribute importance by noting the differences in the $\mu_{jk}(\theta)$ distributions across the ad concepts and ad sizes (within a website at any snapshot in time). In particular, it is clear that the interactions between ad concepts and ad sizes are meaningful: the rank order of the ad concepts' conversion rates varies for different ad sizes. For instance, consider the snapshot of how the TS-HGLM policy evaluated ads and allocated impressions to website $j = 103$ using data through $t = 6$. This is shown as the second row (from the top) of four panels in Figure 2, which we continue to refer to throughout this subsection to explain the findings about ad attributes. For ad size A, the ad concept with the best predicted mean conversion rate is ad concept 4 (14 acquisitions per million impressions), but that same concept is neither the best on the ad size B (mean conversion rate is 131 per million) nor on C (mean conversion rate is 47 per million). In both cases, the best predicted ad concept for sizes B and C is ad concept 3.

The distributions of $\mu_{jk}(\theta)$ shown as boxplots in Figure 2 are the heart of the TS procedure. They represent our current beliefs in the conversion rates reflecting our uncertainty in the HGLM parameters based on the data through t periods. At the end of each period, we simulate G draws from each of those distributions. Using these empirical distributions, we approximate the probability that each ad has the highest mean for each website-by-size pair.

As a result of this procedure, the right side of each panel of Figure 2 shows the next set of allocation probabilities, $w_{j,k,t+1}$, within each ad size, website, and time period for all ad concepts. Looking at these allocation probabilities for $j = 103$ using data through $t = 6$, we

see that for sizes B and C, ad concept 1 is hardly given any impressions in the next period. However, for size A, ad concept 1 is actually predicted to be as good as ad concept 3.

Figure 2 not only shows the importance of attributes (differences within a website across ads), but it also shows learning (changes within an ad-website combination over time) and heterogeneity (differences across websites). The MAB policy learns parameters over time. In our case, it is not practical to report how all parameters are learned, but we highlight how the TS-HGLM policy updates its beliefs about $\mu_{jk}(\theta)$ for particular ad-website combinations. It is clear from Figure 2 that the distributions are wider after the initial period ($t = 1$) than they are after more data has accumulated ($t = 6$).

For instance, after the initial period ($t = 1$) for ad size B and ad concept 3, the predicted distribution of the conversion rate has a 95% interval of (0.92, 56.65) with a mean of 7.35 per million. The probability that it is optimal is 27%. Later on, after the policy learns more about the parameters ($t = 6$), we see that the interval not only shrinks (0.82, 10.31), but also shifts its mean to 2.93 customers per million impressions. This leads the MAB policy to assign a higher probability that this ad concept is optimal, hence allocating 41% of impressions for the next period.

The unobserved heterogeneity in the hierarchical model leads allocations to differ across websites. For example, the two websites in Figure 2 have different winning ads. After $t = 6$ periods, for website $j = 103$, the predicted winners for each ad size (A, B, and C) are ad concepts 4, 3, and 3, whereas those for website $j = 149$ are ad concepts 1, 3, and 4, respectively. Capturing such patterns of website-to-website differences enables the proposed MAB policy to reach greater improvement than other MAB policies that may ignore those patterns.

The key benefit of partial pooling is capturing heterogeneity across websites, but an added benefit is providing a predictive distribution for the ads on any website in question, even in the

absence of a large amount of data on that website. Such sparse data on any one website is a natural feature of this problem. If we were to rely on the observed data alone, especially early in the experiment, we would see that observed conversion rates would be highly misleading. After the initial period for website $j = 149$, there were zero conversions in total, except for some customer acquisition from ad concept 2 on ad size B. That would be rated the best ad concept and ad size combination if we were only using the observed conversion rate for evaluating the ads. But can we really trust that signal given the rare incidence rate in the environment? Trusting that data alone, without leveraging other information, would be problematic; typically, such oversight leads to significant variability in performance of any policy that relies heavily on observed data (e.g., policies referred to as greedy) and independently on each unit's observations (e.g., policies that lack partial pooling across websites). Thus, we leave aside such volatile policies, but in the next section, we examine counterfactual policy simulations (i.e., how a variety of other policies would have performed).

6 Policy Simulations Based on Field Experiment Data

How would a bandit policy perform if we ignored the hierarchical structure, but only accounted for the attribute structure using a homogeneous binomial regression (e.g., TS-GLM)? What if we ignored both the hierarchical and attribute structure, making it a one-factor test treating each action independently, using a binomial model without a regression (e.g., TS-Binomial)? What if we simply stopped the experiment after five periods, and for the remaining time only served the best performing ad(s) at that point (e.g., test-rollout)? This section considers what would have happened if we used other MAB methods in the field experiment. These counterfactual policy simulations reveal which aspects of the method are accounting for improved performance. We begin by detailing how these simulations are constructed.

6.1 Performing Policy Simulations

To run these counterfactual policy simulations, we have to decide on the “truth” (i.e., specify the data-generating process). In particular, we need to set the true conversion rates for each ad on each website. To come up with these conversion rates, we consider two options: a fully model-based approach and a non-parametric approach.

The fully model-based approach uses the exact model (e.g., HGLM) from our proposed MAB policy. This means using the model parameters (mean of distributions) obtained from the actual experimental data through all periods. By construction, this favors the proposed policy because it would generate data from a hierarchical logistic regression model and estimate a hierarchical logistic regression model with TS to show this policy performs best. This can be misleading, yet it is often unquestioned in the practice of evaluating MAB policies (e.g., Filippi et al. 2010; Hauser et al. 2009). Instead of validating MAB policy performance in a realistic setting, this type of policy simulation quantifies how model misspecification (in contrast to the full model assumed to be true) is translated into relative loss of bandit performance.

We therefore utilize a non-parametric approach instead. Like the model-based approach, all of the data across time is included. However, we compute the observed conversion rates (e.g., cumulative conversions divided by cumulative impressions) for each combination of website and ad at the end of the experiment. Those conversion rates are then used as the binomial success rates. In simulation, the conversions (successes) are generated, fixing the number of impressions (trials) to the observed impression count in each decision period for each place (summing across ads), which was already pre-determined by the firm’s media schedule before the experiment. Since we compute conversion rates separately for each ad-website combination, our data-generating process does not assume there is any particular structure in how important ad attributes are or how much websites differ from one another.

Given a true conversion rate, the key assumption is that the truth is a stationary binomial model, so each website-ad combination has a stationary conversion rate. In addition, we assume that the conversion rate of any ad on a website is unaffected by the number of impressions of that ad, that website, or any other ad or website. This assumption is known as the Stable Unit Treatment Value Assumption (SUTVA; Rubin 1990).

We obtain all of the remaining empirical results in this section using this same data-generating process. These true parameters define our MAB problem. In addition to selecting the data-generating process for the policy simulations, we need to decide how we will measure performance. Our main measure of performance is the total number of customers acquired, averaged across replications. We scale this to be the aggregate conversion rate of customers per million impressions. We use a measure commonly seen in industry, which is expected lift above the expected reward earned during the experiment, if the firm were to run a balanced design (the null benchmark). An equal-allocation policy (static experiment with balanced design) earns an average reward equal to the average of the actions, $\frac{1}{K} \sum_k \mu_k(\theta)$. Intuitive to a manager and useful from a practical perspective, lift captures the improvement of any bandit policy over commonly practiced static A/B/n or multivariate tests.

We will discuss a variety of benchmark MAB policies, so we analyze the performance in groups. The first counterfactual simulation we perform is intended to show consistency with the field experiment on which all the subsequent analyses are based.

6.2 Replicating the Field Experiment

In the actual field experiment, we implemented two policies: TS-HGLM and equal allocation. While we observe the adaptive TS-HGLM group improve by 8% over a baseline (as shown in Figure 1), is this difference really meaningful in a statistical sense? We replicate

this experiment via simulation to capture the uncertainty around the observed performance of these two policies, which were actually implemented. This simulation is designed to serve as 100 replications of the field experiment. The resulting simulated worlds allow us to compute predictive distribution of the observed results.

First, we can compare the observed performance of the TS-HGLM (treatment) policy that was actually implemented to the predicted distribution of the balanced design (control) policy. This is analogous to comparing the observed data to a null distribution. We find that the TS-HGLM policy achieves levels of improvement that are outlying with respect to this null distribution, but it takes time for the policy to learn and reach that higher level of performance.

Second, we can compare the full distribution of the balanced design to the full distribution of the TS-HGLM policy. On average, they should reflect the performance observed when they were actually implemented. As expected, these results match the actual relative performance of the two methods: TS-HGLM achieves 8% higher mean performance than equal allocation (4.717 versus 4.373 conversions per million). This consistency gives validity to the counterfactuals to follow. In effect, this shows that our data-generating process and implementation of these two policies can recover the actual performance. The key benefit of looking at simulated versions of the same policies that we implemented in the field experiment is that it enables us to examine the variability in performance and then turn off components of these policies to motivate other policies (which we do for the rest of this section).

To quantify the difference in performance of any pair of policies, we compute a posterior predictive p-value (ppp), the probability that one policy has performance greater than or equal to the performance of another. This is computed empirically using each policy's replicates. Despite each policy's variability in performance across worlds, the TS-HGLM policy out-performs the equal allocation policy in every sampled world ($ppp = 1.00$). This is not very

surprising; the equal-allocation policy is a weak benchmark policy for comparison. Although the balanced design was the firm’s previous plan for running the multivariate test, it is not a strong enough benchmark for fully evaluating MAB policies.

6.3 Evaluation of Benchmark Policies

We evaluate a range of alternative MAB policies to inform our two key decisions: model and allocation rule. First, we consider TS to be the allocation rule, but examine which is the best model to go with it. Then, we consider which is the best-performing randomized allocation rule by comparing TS with standard MAB heuristics from reinforcement learning. Finally, we also consider a managerially relevant and intuitive way to trade off exploration and exploitation: run a balanced test, pick the winning ad, and roll out the winner (i.e., test-rollout policies). For this group of policies, we consider a range of stopping times. Finally, since these are simulations, we can also include the oracle policy for reference, although it is not feasible in practice. The oracle policy always delivers the truly best ad on each website (i.e., as if the oracle knew the aforementioned true data-generating process).

6.3.1 Evaluating the Model Component of the MAB Policy

We now examine a range of MAB policies from complex to simple, all derived from the hierarchical generalized linear model with TS by shutting off the MAB policy’s components one at a time. Figure 3 shows the boxplots for each policy’s distribution of total reward accumulated by the end of the experiment. The results of the TS with the partially-pooled / heterogeneous regression (TS-HGLM); latent-class regression (TS-LCGLM); pooled / homogeneous regression (TS-GLM); and binomial (TS-binomial) policies are all compared to the equal-allocation

policy (Balanced) and the oracle policy (Oracle) in Figure 3, Table 1, and Table 2.⁴ These confirm that the inclusion of continuous heterogeneity (partial pooling via the hierarchical model) is a major driver of performance.

[INSERT FIGURE 3 ABOUT HERE]

[INSERT TABLE 1 ABOUT HERE]

[INSERT TABLE 2 ABOUT HERE]

The results for these TS-based policies suggest that the partial pooling aspect of our proposed policy is important. Recall that the TS-HGLM policy yields an 8% increase in mean above a balanced design. The TS-GLM policy and TS-binomial policy each yields only a 3% improvement above a balanced design. The TS-LCGLM policy falls between those, but only at 4%. The variability and degree of overlap among these policies is shown in Table 2, which shows the probability that any one of these policies using TS performs better than any one of the others. For instance, at the lower end of the performance range, the TS-binomial policy outperforms Balanced policy ($ppp = 0.73$), while at the higher end of the range, the TS-HGLM performs better than its latent-class counterpart, TS-LCGLM ($ppp = 0.60$).

6.3.2 Evaluating the Allocation Rule Component of the MAB Policy

With the evidence clearly pointing in the direction of the HGLM model combined with the TS allocation rule, we now turn to evaluate a range of alternative allocation rules. We consider standard MAB heuristic policies from the reinforcement learning literature (Sutton and Barto 1998) such as greedy and epsilon-greedy.⁵ The greedy policy allocates all observations to the

⁴The model in TS-LCGLM is a logistic regression with two latent classes where all ad attribute parameters vary across the latent classes.

⁵We exclude *upper confidence bound* (UCB) policies despite their coverage in the machine learning and statistical learning literature (Auer 2002; Auer et al. 2002; Filippi et al. 2010; Lai 1987) because these are deterministic solutions. As a consequence, there is no agreed-upon way to transform their indices to allocation probabilities for our batched decision problem. The same is true with other index policies, including the Gittins index.

ad with the largest aggregate observed mean based on all cumulative data. That is, the greedy policy is a myopic and deterministic policy that reflects pure exploitation without exploration. It can change which ad it allocates observations to during each period, but it will always be a “winner-take-all” allocation. The epsilon-greedy policy is a randomized policy that mixes exploitation with a fixed amount of exploration. For any $0 \leq \varepsilon \leq 1$, then ε of observations allocated uniformly across the K ads, and $1 - \varepsilon$ of observations are allocated to the ad with the largest aggregate observed mean (as in the greedy policy). We employ this with with exploration parameter ε set to 10% and 20% (epsgreedy10 and epsgreedy20, respectively). At the extremes, epsilon-greedy nests both a balanced design of equal allocation ($\varepsilon = 100\%$) and a greedy policy ($\varepsilon = 0\%$).

[INSERT FIGURE 4 ABOUT HERE]

[INSERT TABLE 3 ABOUT HERE]

[INSERT TABLE 4 ABOUT HERE]

The greedy policy has a higher mean and more variability than both epsilon-greedy policies (Figure 4 and Table 3). This is expected since ε controls the riskiness of the policy. The epsilon-greedy policies with 10% and 20% perform similarly. The main difference is that $\varepsilon = 20\%$ leads to less variability on the downside of performance, leading to a better worst-case scenario. Again, the pairwise comparisons of the full distributions of performance illustrate that TS-HGLM dominates the greedy and both epsilon-greedy policies ($ppp \geq 0.90$ for all three comparisons).

6.3.3 Evaluating Different Stopping Times for Equal Allocation Policies

Finally, we consider fairly intuitive policies with clear managerial interpretation and call these the test-rollout policies. For a fixed amount of time, the firm runs a balanced design, then

estimates a population-level logit model after the test and allocates all subsequent observations to the ad with the highest-predicted conversion rate. This reflects a complete switch from exploration to exploitation (learn, then earn) as opposed to a simultaneous mixture of the two or a smooth transition from one to the other (earning while learning). At the extreme, when the test lasts all periods, the test-rollout policy reduces a static balanced design.

[INSERT FIGURE 5 ABOUT HERE]

[INSERT TABLE 5 ABOUT HERE]

[INSERT TABLE 6 ABOUT HERE]

We implemented the test-rollout heuristic with six different lengths of the initial period of balanced design (pure exploration). While the average performance for different amounts of initial learning does not change substantially, all achieve approximately a 2% improvement above keeping a balanced design (pure exploration) for all of the periods (Figure 5 and Table 5). Table 6 confirms that TS-HGLM outperforms the whole group of test-rollout policies similarly and consistently ($ppp \geq 0.90$ for all test-period lengths).

The variation within any test-rollout policy is greater than the differences between them. Just looking at mean performance, picking the winner after the initial testing period with a balanced design that lasts for 2 periods yields only a slightly better average performance than when it lasts for 1, 3, 4, 5, or 6 periods. However, Table 6 shows that the probability with which that policy out-performs any other is quite small (between $ppp = 0.51$ and $ppp = 0.64$). So there is not strong evidence of a clear winner among these policies. This may be idiosyncratic to the present context, but it does confirm that such a test-rollout policy is quite sensitive to the choice of the test-period length, which one would not know how to set in practice. In other words, the intuitive test-then-learn approach appears to be much less reliable than an earning while learning MAB policy. Second, the variability in performance is asymmetric. The upper

tails (better conversion rate) of policies' performance distributions do not vary as much as their lower tails. The longer the test period, the smaller the variability around performance because the potential downside is reduced. At the extreme, however, if we consider a balanced test without the rollout phase, the potential upside of the performance distribution would diminish.

7 General Discussion

We have focused on improving the practice of real-time adaptive experiments with online display advertisements to acquire customers. We achieved this by identifying the components of the online advertiser's problem and mapping them onto the existing MAB problem framework. The component missing from existing MAB methods is a way to account for unobserved heterogeneity (e.g., ads differ in effectiveness when they appear on different websites) in the presence of a hierarchical structure (e.g., ads within websites). We extended the existing MAB policies to form a TS-HGLM policy, a natural marriage of hierarchical regression models and randomized allocation rules. In addition to testing this policy against benchmarks in simulation, we implemented it in a live field experiment with a large retail bank. The results were encouraging. We not only demonstrated an 8% increase in customer acquisition rate by using the TS-HGLM policy instead of a balanced design, but we also showed that on average, even strong benchmark MAB policies only reached a level of a 4% increase, half that of the proposed policy.

Nevertheless, there are some limitations to our field experiment and simulations, which may offer promising future directions for research. We acknowledge that acquisition from a display ad is a complex process, and our aim is not to capture all aspects of it. One particular aspect that we do not address is multiple ad exposures. It is natural to imagine the reality that an individual saw more than one of the K ads during the experiment or had multiple exposures to the same

ad. Our data does not contain individual-level (or cookie-level) information, but this could be an interesting area of research, trying to combine ad attribution models with MAB policies. The issue of multiple exposures would raise concerns in this paper if the following conditions were true: (i) the repeated viewing of particular types of ads has a substantially different impact on acquisition than the repeated viewing of other types of ads; (ii) that difference is so large that a model including and a model ignoring repeated exposures each identifies a different winning ad; and (iii) there is a difference in the identified winning ad for many of the websites with large impression volume. While this scenario is possible, we argue it is unlikely. As further evidence, we also do not see those problematic time dynamics of conversion rates at the website-ad level. The data suggest the assumption of a constant μ_{jk} is reasonable during our experiment. Considering the short time window in which display advertising campaigns run, this assumption is not particularly limiting.

Another limitation of our work is that we do not take into account the known and finite time horizon. While we used 10 periods, we did not take this time horizon into account while making allocations for periods 1 through 9. In a typical dynamic programming solution, one considers either backward induction from the end point or some other explicitly forward-looking recursion. If the relative cost required to run the experiment is negligible, then there is little gained from optimizing the experiment during that period. In fact, this reduces to a test-rollout setting where it is best to learn then earn. By contrast, if the observations are relatively costly or if there is always earning to be gained from learning (and such learning takes a long time), then it would be useful to consider a MAB experiment for an infinite horizon. However, most MAB experiments fall somewhere between those two extremes. Perhaps the length of the MAB experiment is a decision that the experimenter should optimize. This extra optimal stopping problem is the focus of a family of methods known as expected value of information gained

and knowledge gradient methods (Chick and Inoue 2001; Chick and Gans 2009; Powell 2011).

While we have utilized the fact that batch size was exogenous and given to us for each website and each period (M_{jt}), we could generalize our problem to a setting where we had control of the batch size and were making allocations of impression volume across websites. This is relevant as real-time bidding for media on ad exchanges becomes even more common. However, introduces complexities to the MAB problem, such as correlations among impression volume, cost per impression, and expected conversion rate. In addition, we would need to consider methods that explicitly consider the batch size (Chick and Gans 2009; Chick et al. 2010; Frazier et al. 2009).

One key limitation comes from our data: We only observe conversion without linking those customers to their subsequent post-acquisition behavior. It seems natural to acquire customers by considering the relative values of their expected customer lifetime value (CLV) and acquisition cost instead of merely seeking to increase acquisition rate (i.e., lower cost per acquisition). Sequentially allocating resources to acquire customers based on predictions about their future value seems like a promising marriage between MAB and CLV.

Finally, we see the bandit problem as a powerful framework for optimizing a wide range of business operations. This broader class of problems is centered around the question, “Which targeted marketing action should we take, when, with which customers, and in which contexts?” As we continue equipping managers and marketing researchers with these tools to employ in a wide range of settings, we should have a more systematic understanding of the robustness and sensitivity of these methods to common practical issues.

References

- Agarwal, Deepak, Bee-Chung Chen, Pradheep Elango. 2008. Explore/Exploit Schemes for Web Content Optimization. *Yahoo Research paper series* .
- Agrawal, Shipra, Navin Goyal. 2012. Analysis of Thompson Sampling for the multi-armed bandit problem. *Journal of Machine Learning Research Conference Proceedings, Conference on Learning Theory* **23**(39) 1–26.
- Anderson, Eric, Duncan Simester. 2011. A Step-by-Step Guide to Smart Business Experiments. *Harvard Business Review* **89**(3) 98–105.
- Auer, Peter. 2002. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research* (3) 397–422.
- Auer, Peter, Nicolo Cesa-Bianchi, Paul Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning* **47** 235–256.
- Bates, Douglas, Martin Maechler, Ben Bolker, Steven Walker. 2013. R Package 'lme4' [Http://cran.r-project.org/web/packages/lme4/lme4.pdf](http://cran.r-project.org/web/packages/lme4/lme4.pdf).
- Bates, Douglas, Donald G. Watts. 1988. *Nonlinear Regression Analysis and Its Applications*. Wiley.
- Berry, Donald A. 1972. A Bernoulli Two-Armed Bandit. *Annals of Mathematical Statistics* **43** 871–897.
- Berry, Donald A., Bert Fristedt. 1985. *Bandit Problems*. Chapman Hall.
- Bertsimas, Dimitris, Adam J. Mersereau. 2007. Learning Approach for Interactive Marketing. *Operations Research* **55**(6) 1120–1135.
- Bradt, R. N., S. M. Johnson, S. Karlin. 1956. On Sequential Designs for Maximizing the Sum of n Observations. *Annals of Mathematical Statistics* **27**(4) 1060–1074.
- Chapelle, Olivier, Lihong Li. 2011. An Empirical Evaluation of Thompson Sampling. *Online Trading of Exploration and Exploitation workshop* 1–6.
- Chick, S.E., N. Gans. 2009. Economic analysis of simulation selection problems. *Management Science* **55**(3) 421–437.
- Chick, S.E., K. Inoue. 2001. New Two-Stage and Sequential Procedures for Selecting the Best Simulated System. *Operations Research* **49**(5) 732–743.
- Chick, S.E., Branke J., Schmidt C. 2010. Sequential Sampling to Myopically Maximize the Expected Value of Information. *INFORMS Journal on Computing* **22**(1) 71–80.
- Dani, V., T. P. Hayes, S. M. Kakade. 2008. Stochastic Linear Optimization Under Bandit Feedback. *Conference on Learning Theory* .
- Davenport, Thomas H. 2009. How to Design Smart Business Experiments. *Harvard Business Review* **87**(2) 1–9.

- Donahoe, John. 2011. How ebay Developed a Culture of Experimentation: HBR Interview of John Donahoe. *Havard Business Review* **89**(3) 92–97.
- eMarketer. 2012a. Brand Marketers Cling Direct Response Habits Online. Website. www.emarketer.com/Article/Brand-Marketers-Cling-Direct-Response-Habits-Online/1008857/ Accessed 30 Mar 2013.
- eMarketer. 2012b. Digital Ad Spending Tops \$37 billion. Website. www.emarketer.com/newsroom/index.php/digital-ad-spending-top-37-billion-2012-market-consolidates/ Accessed 30 Mar 2013.
- Filippi, Sarah, Olivier Cappé, Aurélien Garivier, Csaba Szepesvári. 2010. Parametric bandits: The generalized linear case. J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, A. Culotta, eds., *Advances in Neural Information Processing Systems* **23**. 586–594.
- Frazier, P.I., W.B. Powell, S. Dayanik. 2009. The Knowledge-Gradient Policy for Correlated Normal Beliefs. *INFORMS Journal on Computing* **21**(4) 599–613.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, Donald B. Rubin. 2004. *Bayesian Data Analysis*. 2nd ed. Chapman & Hall, New York, NY.
- Gelman, Andrew, Jennifer Hill. 2007. *Data Analysis Using Regression and Multi-level/Hierarchical Models*. Cambridge University Press, New York, NY.
- Gittins, John C. 1979. Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society, Series B* **41**(2) 148–177.
- Gittins, John C., Kevin Glazebrook, Richard Weber. 2011. *Multi-Armed Bandit Allocation Indices*. 2nd ed. John Wiley and Sons, New York, NY.
- Gittins, John C., D. M. Jones. 1974. A dynamic allocation index for the sequential design of experiments. J. Gani, K. Sarkadi, I. Vineze, eds., *Progress in Statistics*. North-Holland Publishing Company, Amsterdam, 241–266.
- Granmo, O.-C. 2010. Solving Two-Armed Bernoulli Bandit Problems Using a Bayesian Learning Automaton. *International Journal of Intelligent Computing and Cybernetics* **3**(2) 207–232.
- Hauser, John R., Glen L. Urban, Guilherme Liberali, Michael Braun. 2009. Website Morphing. *Marketing Science* **28**(2) 202–223.
- Kaufmann, Emilie, Nathaniel Korda, Remi Munos. 2012. Thompson Sampling: An Asymptotically Optimal Finite Time Analysis <http://arxiv.org/abs/1205.4217/>.
- Krishnamurthy, V., Bo Wahlberg. 2009. Partially Observed Markov Decision Process Multi-armed Bandits: Structural Results. *Mathematics of Operations Research* **34**(2) 287–302.
- Lai, T. L. 1987. Adaptive Treatment Allocation and the Multi-Armed Bandit Problem. *Annals of Statistics* **15**(3) 1091–1114.
- Lin, Song, Juanjuan Zhang, John R. Hauser. 2013. Learning from Experience, Simply. *Working Paper* .

- Manchanda, Puneet, Jean-Pierre Dubé, K.Y. Goh, P.K. Chintagunta. 2006. The Effect of Banner Advertising on Internet Purchasing. *Journal of Marketing Research* **43**(February) 98–108.
- May, Benedict C., Nathan Korda, Anthony Lee, David S. Leslie. 2011. Optimistic Bayesian Sampling in Contextual Bandit Problems. *Department of Mathematics, University of Bristol* (Technical Report 11:01).
- Meyer, Robert J., Y. Shi. 1995. Sequential Choice Under Ambiguity: Intuitive Solutions to the Armed-Bandit Problem. *Management Science* **41**(5) 817–834.
- Ortega, Pedro A., Daniel A. Braun. 2010. A Minimum Relative Entropy Principle for Learning and Acting. *Journal of Artificial Intelligence Research* **38** 475–511.
- Ortega, Pedro A., Daniel A. Braun. 2013. Generalized Thompson Sampling for Sequential Decision-Making and Causal Inference <http://arxiv.org/abs/1303.4431>.
- Osband, Ian, Daniel Russo, Benjamin Van Roy. 2013. (More) Efficient Reinforcement Learning via Posterior Sampling <http://arxiv.org/abs/1306.0940>.
- Powell, Warren B. 2011. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley, New Jersey.
- Reiley, David, Randall Aaron Lewis, Panagiotis Papadimitriou, Hector Garcia-Molina, Prabhakar Krishnamurthy. 2011. Display Advertising Impact: Search Lift and Social Influence. *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1019–1027.
- Robbins, H. 1952. Some Aspects of the Sequential Design of Experiments. *Bulletin of the American Mathematics Society* **58**(5) 527–535.
- Rubin, Donald. 1990. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* **66**(5) 688–701.
- Rusmevichientong, Paat, John N. Tsitsiklis. 2010. Linearly Parameterized Bandits. *Mathematics of Operations Research* **35**(2) 395–411.
- Russo, Daniel, Benjamin Van Roy. 2013. Learning to Optimize Via Posterior Sampling <http://arxiv.org/abs/1301.2609>.
- Scott, Steven L. 2010. A Modern Bayesian Look at the Multi-Armed Bandit. *Applied Stochastic Models Business and Industry* **26**(6) 639–658.
- Sutton, Richard S., Andrew G. Barto. 1998. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.
- Thompson, Walter R. 1933. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* **25**(3) 285–294.
- Tsitsiklis, John N. 1986. A Lemma on the Multi-Armed Bandit Problem. *IEEE Transactions on Automatic Control* **31**(6) 576–577.
- Urban, Glen L., Guilherme Liberali, Erin MacDonald, Robert Bordley, John R. Hauser. 2013. Morphing Banner Advertising. *Marketing Science*, forthcoming .

Wahrenberger, David L., Charles E. Antle, Lawrence A. Klimko. 1977. Bayesian Rules for the Two-armed Bandit Problem. *Biometrika* **64**(1) 1724.

White, John Myles. 2012. *Bandit Algorithms for Website Optimization*. O'Reilly Media.

Whittle, P. 1980. Multi-armed Bandits and the Gittins Index. *Journal of Royal Statistical Society, Series B* **42**(2) 143–149.

Appendix A: Values from Figure 2

Tables 7, 8, and 9 provide the underlying key values illustrated in the panels of Figure 2 (one table for each ad size), as well as the observed data of cumulative conversions and impressions broken down by two time periods ($t = 1$ and 6), two websites ($j = 103$ and 149), each ad size (A, B, and C), and each ad concept (1, 2, 3, and 4). The belief distributions of $\mu_{jk}(\theta)$ for all k and the two j are summarized by mean and 95% intervals. The resulting allocations, $w_{j,k,t+1}$, are shown in the tables and match those shown in Figure 2.

[INSERT TABLE 7 ABOUT HERE]

[INSERT TABLE 8 ABOUT HERE]

[INSERT TABLE 9 ABOUT HERE]

8 Tables and Figures

	Mean	SD	Improvement above Balanced	Efficiency to Oracle	Improvement Balanced to Oracle	Relative Mean	Relative Precision
Oracle	5.932	0.078	36%	100%	100%	26%	33%
TS-HGLM	4.717	0.090	8%	80%	22%	0%	0%
TS-LCGLM	4.527	0.088	4%	76%	10%	-4%	4%
TS-GLM	4.493	0.091	3%	76%	8%	-5%	-3%
TS-binomial	4.493	0.087	3%	76%	8%	-5%	5%
Balanced	4.373	0.138	0%	74%	0%	-7%	-58%

Table 1: *Summary of performance for MAB policies using TS.* The policy TS-HGLM is the best performing policy, and it is compared to other TS-based policies. Balanced refers to the equal allocation policy, and Oracle refers to the hypothetical ideal policy. They are shown in subsequent tables to show the lower and upper end of policies.

		Bal	T-B	T-G	T-LG	T-HG
Balanced	Bal	1				
TS-binomial	T-B	0.73	1			
TS-GLM	T-G	0.78	0.44	1		
TS-LCGLM	T-LG	0.83	0.61	0.60	1	
TS-HGLM	T-HG	1.00	0.96	0.95	0.96	1

Table 2: *Pairwise comparison of policies using TS.* The values are the probability that the policy in the row is better than (or at least as good as) the policy in the column. For instance, there is a 96% probability that the TS-HGLM policy achieves at least as high of an aggregate conversion rate as the TS-GLM policy. These posterior predictive p-values were computed via simulation from the distribution of aggregate conversions shown in boxplots. The same method is used throughout the tables with pairwise comparisons of policies.

	Mean	SD	Mean Improvement above Balanced	Mean Efficiency to Oracle	Mean Improvement Balanced to Oracle	Relative Mean	Relative Precision
Oracle	5.932	0.078	36%	100%	100%	26%	33%
TS-HGLM	4.717	0.090	8%	80%	22%	0%	0%
greedy	4.520	0.115	3%	76%	9%	-4%	-39%
epsgreedy(10)	4.489	0.094	3%	76%	7%	-5%	-9%
epsgreedy(20)	4.504	0.089	3%	76%	8%	-5%	2%
Balanced	4.373	0.138	0%	74%	0%	-7%	-58%

Table 3: *Summary of performance for greedy and epsilon-greedy policies.* The greedy and epsilon-greedy policies are compared alongside the TS-HGLM, Balanced, and Oracle policies. The policies labeled “epsgreedy(10)” and “epsgreedy(20)” refer to epsilon-greedy policies with the exploration variable ε set to 10% and 20%, respectively.

		Bal	G	EG10	EG20	T-HG
Balanced	Bal	1				
greedy	G	0.79	1			
epsgreedy10	EG10	0.77	0.44	1		
epsgreedy20	EG20	0.78	0.44	0.51	1	
TS-HGLM	T-HG	1.00	0.93	0.97	0.93	1

Table 4: *Pairwise comparison of policies using greedy and epsgreedy.* The two epsgreedy policies have similar performance, since one is better than the other only 51% of the time.

	Mean	SD	Mean Improvement above Balanced	Mean Efficiency to Oracle	Mean Improvement Balanced to Oracle	Relative Mean	Relative Precision
Oracle	5.932	0.078	36%	100%	100%	26%	33%
TS-HGLM	4.717	0.090	8%	80%	22%	0%	0%
testrollout-t6	4.450	0.085	2%	75%	5%	-6%	10%
testrollout-t5	4.446	0.098	2%	75%	5%	-6%	-16%
testrollout-t4	4.463	0.099	2%	75%	6%	-5%	-18%
testrollout-t3	4.463	0.108	2%	75%	6%	-5%	-32%
testrollout-t2	4.479	0.128	2%	76%	7%	-5%	-51%
testrollout-t1	4.453	0.155	2%	75%	5%	-6%	-67%
Balanced	4.373	0.138	0%	74%	0%	-7%	-58%

Table 5: *Summary of performance for test-rollout policies.* The test-rollout policies with various lengths of the initial test period are compared alongside the TS-HGLM, Balanced, and Oracle policies. The test-rollout policy with two initial test periods performs better than other test-rollout policies in this setting. The Balanced policy is equivalent to a test-rollout policy with 10 initial test periods without any rollout period afterward.

		Bal	TR1	TR2	TR3	TR4	TR5	TR6	T-HG
Balanced	Bal	1							
testrollout-t1	TR1	0.68	1						
testrollout-t2	TR2	0.74	0.49	1					
testrollout-t3	TR3	0.70	0.54	0.46	1				
testrollout-t4	TR4	0.66	0.51	0.47	0.51	1			
testrollout-t5	TR5	0.64	0.48	0.36	0.48	0.44	1		
testrollout-t6	TR6	0.72	0.48	0.39	0.40	0.46	0.51	1	
TS-HGLM	T-HG	1.00	0.93	0.94	0.97	0.99	1.00	1.00	1

Table 6: *Pairwise comparison of policies using test-rollout policies with different lengths of the initial test period.* Using a two test periods is better than one test period 51% of the time and better than three test periods with a probability of 53%.

size A										
website	time	concept	$w_{j,k,t+1}$	μ_{jk}	Mean	μ_{jk} 2.5%	μ_{jk} 97.5%	y_{jkt}	m_{jkt}	
j103	1	1	0.30	4.76	0.42	43.70	0	13086		
		2	0.27	3.99	0.36	46.47	0	13086		
		3	0.19	2.81	0.19	40.03	0	13086		
		4	0.23	3.64	0.25	43.96	0	13086		
	6	1	0.24	12.99	4.52	35.13	1	96415		
		2	0.17	9.94	2.71	36.93	1	78776		
		3	0.26	12.73	3.84	44.13	1	86540		
		4	0.33	13.88	4.23	41.37	2	97296		
j149	1	1	0.27	5.83	0.55	64.33	0	3572		
		2	0.25	4.84	0.34	64.83	0	3572		
		3	0.22	3.99	0.19	79.68	0	3572		
		4	0.27	4.70	0.29	85.16	0	3572		
	6	1	0.30	6.08	1.09	33.82	1	48028		
		2	0.16	3.43	0.47	22.71	0	38914		
		3	0.28	5.61	0.82	37.47	0	40281		
		4	0.27	5.05	0.66	37.00	0	48360		

Table 7: Values from Figure 2 for ad size A. The predictive distribution of each μ_{jk} based on the model and data through t periods, is summarized by its mean (column labeled “ μ_{jk} Mean”) and 95% interval (columns labeled μ_{jk} 2.5% and μ_{jk} 97.5%). The predictive distributions are based on the actual cumulative number of conversions and impressions (columns labeled y_{jkt} and m_{jkt} , respectively). The subsequent allocation weights are for period $t + 1$ (column labeled $w_{j,k,t+1}$). The above descriptions apply here and to Tables 8 and 9.

size B									
website	time	concept	$w_{j,k,t+1}$	μ_{jk}	Mean	μ_{jk} 2.5%	μ_{jk} 97.5%	y_{jkt}	m_{jkt}
j103	1	1	0.01	76.05	28.40	210.40	1	18215	
		2	0.22	165.29	56.90	492.54	5	18215	
		3	0.41	210.07	78.38	662.32	5	18215	
		4	0.37	206.97	75.22	554.49	3	18215	
	6	1	0.01	88.70	44.36	171.69	2	24814	
		2	0.30	147.29	71.24	303.86	14	88826	
		3	0.52	167.57	89.55	299.38	36	207258	
		4	0.18	131.02	61.53	294.88	8	61298	
j149	1	1	0.16	5.69	1.07	36.85	0	28356	
		2	0.33	9.03	1.41	63.44	1	28356	
		3	0.27	7.35	0.92	56.65	0	28356	
		4	0.23	6.81	1.03	56.76	0	28356	
	6	1	0.21	2.35	0.76	7.41	0	295132	
		2	0.22	2.17	0.59	8.67	1	404384	
		3	0.41	2.93	0.82	10.31	2	403467	
		4	0.15	1.87	0.52	7.18	0	302950	

Table 8: Values from Figure 2 for ad size B.

size C										
website	time	concept	$w_{j,k,t+1}$	μ_{jk}	Mean	μ_{jk} 2.5%	μ_{jk} 97.5%	y_{jkt}	m_{jkt}	
j103	1	1	0.10		27.48	6.68	116.32	2	17439	
		2	0.08		21.36	4.27	93.26	1	17439	
		3	0.26		40.23	8.58	190.65	0	17439	
		4	0.57		61.37	14.87	256.00	1	17439	
	6	1	0.02		26.28	12.08	58.81	3	43323	
		2	0.31		45.15	17.26	119.30	4	40787	
		3	0.39		50.49	20.97	121.70	5	102441	
		4	0.28		47.01	19.73	111.75	3	115023	
j149	1	1	0.23		3.31	0.29	31.32	0	14059	
		2	0.18		2.78	0.29	31.35	0	14059	
		3	0.22		2.98	0.20	42.94	0	14059	
		4	0.37		5.01	0.50	75.27	0	14059	
	6	1	0.25		1.63	0.36	6.97	0	186382	
		2	0.20		1.34	0.23	8.13	0	157923	
		3	0.22		1.53	0.32	6.60	0	222576	
		4	0.33		1.90	0.33	9.14	1	288744	

Table 9: Values from Figure 2 for ad size C.

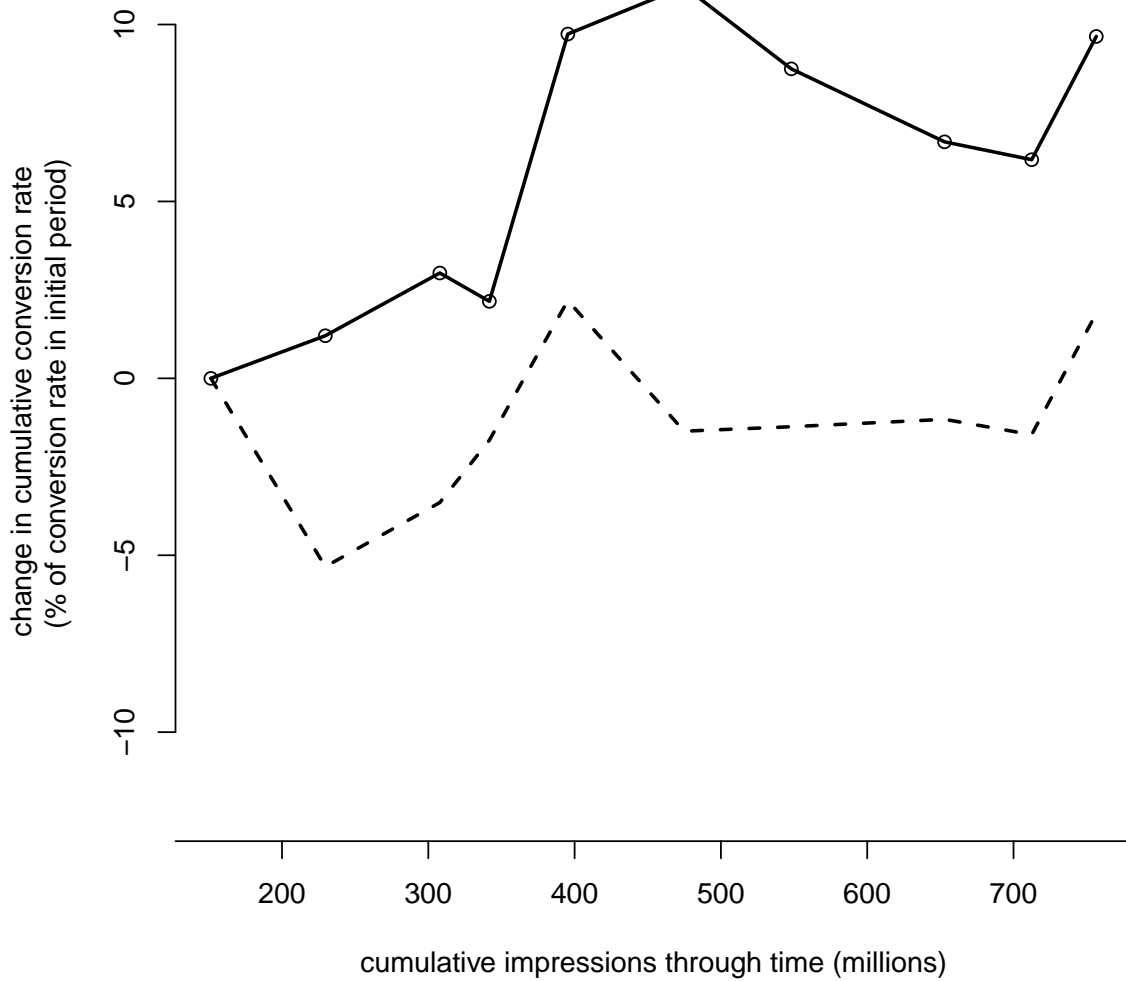


Figure 1: *Results observed in field experiment.* The field experiment results show the TS-HGLM (adaptive group, solid line) achieves a higher cumulative improvement than the balanced design (static group, dashed line), relative to the cumulative conversion rate after the initial period. For the adaptive policy, the circles indicate when reallocations occurred (every five to seven days).

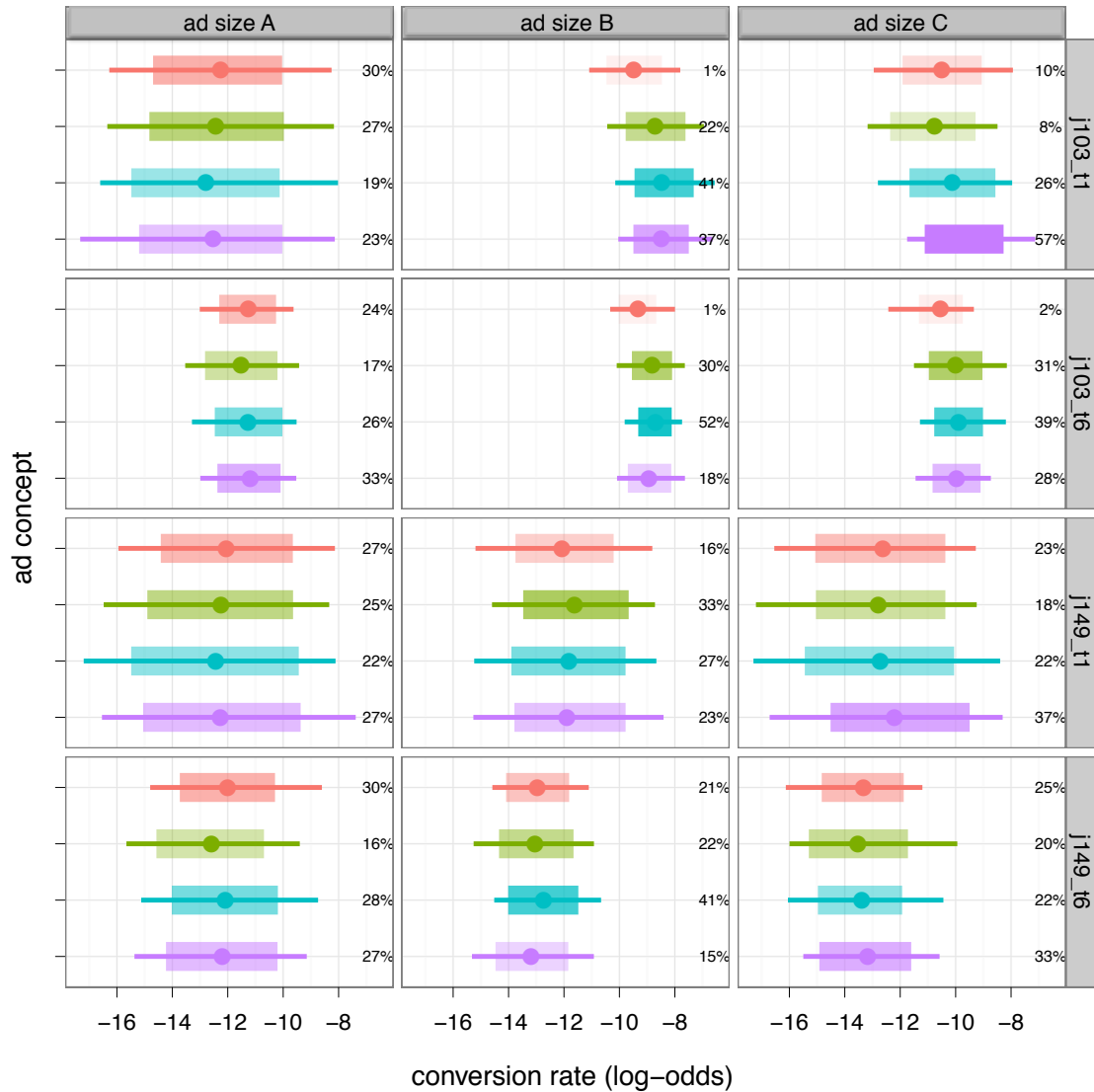


Figure 2: *Heterogeneity in conversion rates across websites and learning over time.* The lines represent the belief distributions of conversion rates, based on predictive distributions of parameters from the HGLM (hierarchical logit model). Within each panel of a website j , time period t , and an ad size, there are four ad concepts (horizontal lines, ordered from top to bottom, ad concepts 1 to 4). The allocation probabilities based on that model are printed (and shown by level of transparency of shading, from invisible 0% to opaque 100%). The four vertical panels show two different websites at two different time periods. Heterogeneity is shown through differences across the two websites (j) for the same time period. Learning is shown through the two time periods (t) for the same website. Both heterogeneity and learning cause allocations to differ across ads. The three panels in each row show the different ad sizes. Appendix A contains tables with the underlying values from this figure ($w_{j,k,t+1}$ and summaries of the posterior distributions of μ_{jk}), as well as the observed data of cumulative conversions and impressions broken down by each website ($j = 103$ and 149), ad size, ad concept, and time period ($t = 1$ and 6).

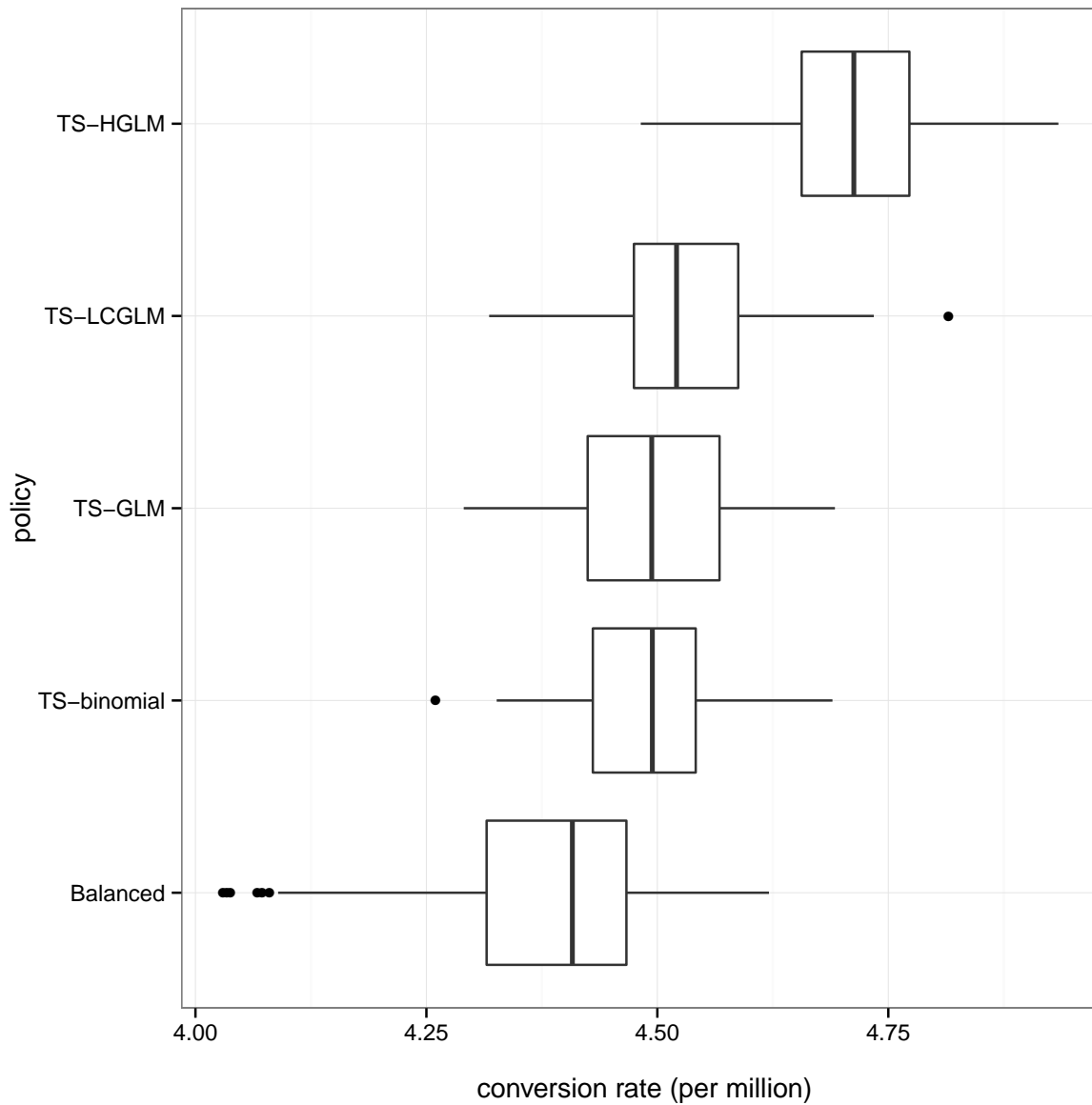


Figure 3: *Distributions of conversions following TS-based policies.* The distributions of total conversions for TS-based are compared. TS-HGLM performs better than the other versions of TS with alternative model specifications, suggesting that the continuous parameter heterogeneity across websites drives the improvement in performance. For all of the boxplots, the center line is the median, the box represents the interquartile range (IQR), the whiskers stretch to $1.5 \times \text{IQR}$ beyond the edges of the IQR, and the points are any values beyond the range of the whiskers.

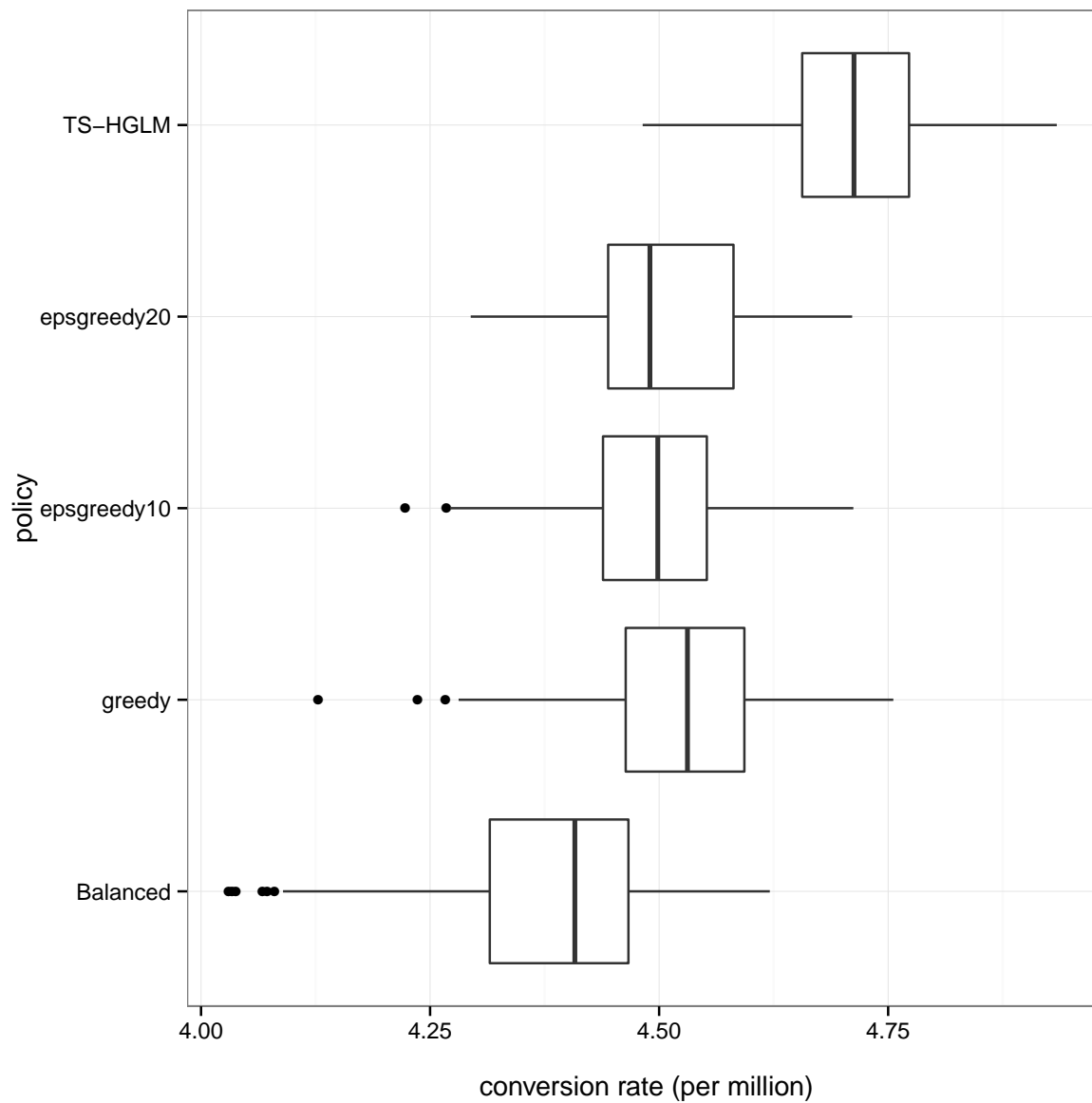


Figure 4: *Distributions of conversions greedy and epsilon-greedy policies.* The distributions of total conversions for greedy and epsilon-greedy policies are compared to the TS-HGLM policy and the balanced design policy. Setting epsilon to 20% performs better than setting it to 10%.

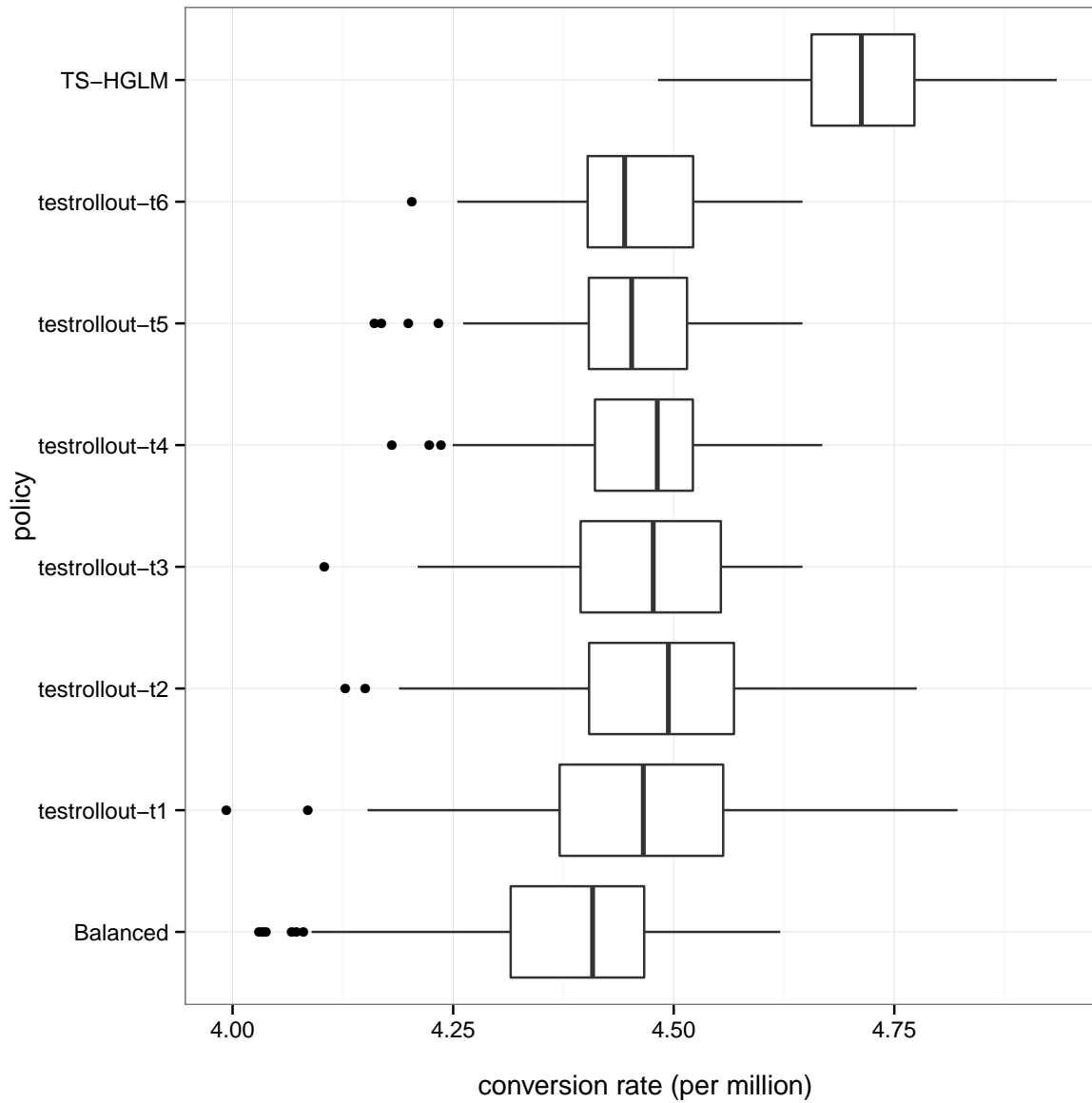


Figure 5: *Distributions of conversions following test-rollout policies.* The distributions of total conversions for test-rollout policies are compared to the TS-HGLM policy and the balanced design policy. Testing for only two initial periods yields better performance than testing for any other length of time between 1 and 6 periods.