AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Adaptive Experimental Design: Prospects and Applications in Political Science

Molly Offer-Westort Stanford Graduate School of Business
Alexander Coppock Yale University
Donald P. Green Columbia University

Abstract: *Experimental researchers in political science frequently face the problem of inferring which of several treatment arms is most effective. They may also seek to estimate mean outcomes under that arm, construct confidence intervals, and test hypotheses. Ordinarily, multiarm trials conducted using static designs assign participants to each arm with fixed probabilities. However, a growing statistical literature suggests that adaptive experimental designs that dynamically allocate larger assignment probabilities to more promising treatments are better equipped to discover the best performing arm. Using simulations and empirical applications, we explore the conditions under which such designs hasten the discovery of superior treatments and improve the precision with which their effects are estimated. Recognizing that many scholars seek to assess performance relative to a control condition, we also develop and implement a novel adaptive algorithm that seeks to maximize the precision with which the largest treatment effect is estimated .*

Verification Materials: The data and materials required to verify the computational reproducibility of the results, procedures, and analyses in this article are available on the *American Journal of Political Science Dataverse* within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/CMUHBU.

Experimentation in the social sciences often involves a search for the intervention that maximizes a desired outcome or yields the largest treatment effect relative to a baseline control condition. Which of the many ways of monitoring corruption among public officials minimizes the amount of missing public funds (Olken 2007)? What combination of personal attributes makes an applicant for naturalization most attractive to voters in the receiving country (Hainmueller and Hangartner 2013)? In many cases, this search dovetails with other academic objectives, such as discerning the causal mechanisms that make certain interventions especially effective (Ludwig, Kling, and Mullainathan 2011).

Experiments that assess the relative effectiveness of competing interventions, be they policies or messages, often confront a fundamental problem: the list of interventions under consideration is so long that it is prohibitively costly and time consuming to sufficiently test the full range of treatment arms. Furthermore, even if money were no object, a prolonged search for the best alternative may impose excessive costs on human subjects and delay the implementation of interventions that would be superior to the status quo. Researchers typically use theory, intuition, and prior research to reduce the large space of possible treatments to a manageable few. Persistent concerns remain, however, that some possibilities are discarded too soon or that practical constraints on the number of arms in the typical experiment induce researchers to narrow the range of interventions they test.

Response-adaptive trials may increase the speed and efficiency with which multiarm trials discern the best performing intervention or interventions. In contrast to conventional static designs that allocate a fixed proportion of subjects to each arm throughout the trial, this class of adaptive designs dynamically update assignment

Molly Offer-Westort is a post-doctoral fellow at the Stanford Graduate School of Business, 655 Knight Way, Stanford, CA 94305 (mollyow@stanford.edu). Alexander Coppock is Assistant Professor of Political Science, Yale University, 115 Prospect Street, New Haven, CT 06511 (alex.coppock@yale.edu). Donald P. Green is the J. W. Burgess Professor of Political Science, Columbia University, 420 W. 118th Street New York, NY 10027 (dpg2110@columbia.edu).

The authors thank Peter Aronow, Fredrik Sävje, and participants of the Stanford Causal Inference Group for valuable comments. We thank the Graduate School of Arts and Sciences at Columbia for support. Studies 2 and 3 were preregistered with Evidence in Governance and Politics (20191121AA and 20181127AC).

American Journal of Political Science, Vol. 00, No. 0, January 2021, Pp. 1–19

DOI: 10.1111/ajps.12597

probabilities based on observed outcomes, investing an ever-larger share of the subject pool in more promising treatment arms. In substantive domains from advertising (Graepel et al. 2010; Li et al. 2010) to biomedical research (Chow and Chang 2008; Chin 2016; Lei, Tewari and Murphy 2017; Villar, Bowden, and Wason 2015; Sydes et al. 2012), adaptive trials are used to speed the search for the best performing intervention.

That said, adaptive designs are no panacea. In situations where several treatment arms are effective, adaptive algorithms may equivocate, allocating more subjects to arms whose initial success was due only to sampling fluctuation. In the case where outcomes under all arms are equivalent, every arm is the "best" arm, and adaptive design confers no special advantages and might even cause a trial to drag on in a vain search for a best arm. Given this uneasy combination of upside potential and downside risk, the literature on adaptive designs abounds with proposals for allocating subjects in ways that guard against false positives and give early warning signals about futile searches among roughly equally effective (or ineffective) interventions—and some algorithms are more robust to recover from such false starts than others (Russo et al. 2017; Urteaga and Wiggins 2017). Alternatives, such as the empirical Bayes Stein-type estimators proposed by Dimmery, Bakshy, and Sekhon (2019), exploit a *set* of best arms, rather than a unique best arm, allowing researchers to better distinguish among the top treatments.

Adaptive trials encompass a broad class of designs that potentially evolve based on interim results. Here, we discuss trials that dynamically update treatment assignment probabilities based on observed outcomes (sometimes referred to as response-adaptive randomization). Other adaptive design adjustments include modifying treatments, updating sample size or eligibility criteria, or halting the trial entirely (Chin 2016; Pallmann et al. 2018; Wason, Brocklehurst, and Yap 2019). Adaptive designs also vary in their goals. Some aim to maximize cumulative response (e.g., in a medical trial, achieving optimal health outcomes across all subjects in the study, as in Murphy 2003); others endeavor to identify arms with average responses that exceed some threshold (Locatelli, Gutzeit, and Carpentier 2016). These alternative objectives fall outside the scope of this article, which covers adaptive trials with objectives that are most relevant to political scientists: finding and evaluating the best performing arm and estimating its causal effect relative to some control condition. For the remainder of the article, we refer to response-adaptive designs simply as "adaptive designs," but we acknowledge that the term "adaptive" encompasses a much broader class of designs.

The first aim of this article is to introduce political scientists to such adaptive designs, suggest approaches for estimation in such settings, and highlight the conditions under which adaptive designs outperform conventional static designs, focusing on the metrics of best arm selection and root mean squared error (RMSE). We begin by introducing a commonly used algorithm for response-adaptive randomization, known in the literature as Thompson sampling (Thompson 1933, 1935). Although alternative algorithms may outperform standard Thompson sampling for a given objective, this procedure is an apt starting point. It is one of the simplest, most intuitive, and most widely studied approaches to response-adaptive randomization, and versions of Thompson sampling are extensively used in industry applications.

The second aim of this article is to propose a new adaptive algorithm specifically tailored to a common research goal for many political scientists: estimating the average treatment effect of the best performing arm (relative to a control condition) with as much precision as possible. Our *control-augmented* algorithm adaptively allocates more subjects to the best performing arm as it emerges (as in standard Thompson sampling) but also allocates more subjects to the control arm as the set of effective treatments is whittled down.

To see the intuition behind the control-augmented design, consider a much-simplified adaptive algorithm. The full sample is divided into two equally sized batches. With the first batch, we conduct a large pilot study in which we randomly assign subjects to each of the treatment arms with equal probabilities. With the second batch, we randomly assign subjects to one of just two arms: the control or the best performing arm from the first batch. Because such a large fraction of the total sample is allocated to the control and best-performing arm by the end of the study, we achieve large increases in precision over a standard static trial. This simplified algorithm divides the full sample into just two large batches, whereas in typical applications, the control-augmented design smooths the process of allocating additional subjects to the control and the best-performing arm over many smaller batches, thereby using the full sample more efficiently.[1]

We illustrate the features of adaptive designs with a series of simulations in more and less favorable scenarios.

[1]That said, we find in simulations in Supporting Information (SI) D.1 (p. 21) that the largest precision increases occur when moving from one batch (a static trial) to two batches. Researchers wishing to avoid the logistical and analytic complications of a full-blown adaptive algorithm might find this two-step procedure to be an attractive alternative to a standard static multiarm design.

The simulation results can help guide researchers who want to learn about the conditions under which adaptive designs improve precision. We then turn to a set of empirical applications. Study 1 uses Thompson sampling to measure support for alternative ballot measures for two policies: increasing the minimum wage and a right-to-work law. In the case of right-to-work proposals, the adaptive design quickly identifies a clear winner with a high degree of statistical precision. Results are more ambiguous for minimum wage proposals, where several possibilities seem equally promising. Study 2 applies our control-augmented algorithm to the study of misperceptions of facts. We explore six alternative treatments to induce survey respondents to give more correct answers to factual questions about economic conditions. We conduct our adaptive trial separately among Republican and Democratic partisans; among Republicans, the best treatment arm (being entered in a lottery for an Amazon gift card) prevailed quickly, whereas among Democrats, the lottery took longer to beat out competitor arms. We offer an additional application for adaptive designs in SI G (p. 39): We apply a model-based adaptive algorithm to the factorial conjoint experimental designs that are increasingly used in political science. This approach allows us to navigate the large number of combinations of ballot measure attribute levels, with the objective of finding the combination that will yield the highest support.

Adaptive Trials and the Multiarmed Bandit Problem

Response-adaptive randomized trials are frequently positioned in the framework of the multiarmed bandit problem, first posed by Thompson (1933, 1935), where the experimenter is tasked with sequentially allocating finite resources across multiple treatment arms.[2] Each treatment arm is associated with a distribution of outcomes in the study population.

These distributions are not known at the outset of the experiment, but the experimenter gradually learns about them by observing outcomes under different treatments. The typical trade-off addressed in such settings is between exploration and exploitation. Experimenters would like to *explore* by obtaining information about each arm so that they can be confident in selecting the best arm. They would also like to *exploit* the best performing arms by allocating large proportions of subjects to them in order to achieve the goals of the intervention. These two objectives are in tension with one another. On the one hand, too much exploration means wasting draws on underperforming arms. On the other hand, overexploitation of early frontrunners risks ignoring potentially superior arms.

A common objective in these settings is the maximization of expected reward. In the binary case, reward may be as simple as observed "successes" under each arm, where a success is a specified outcome value such as registering to vote or making a charitable donation. Despite the "success" label, the outcome need not be normatively desirable—the reward function can just as easily be inverted to minimize realizations of the specified outcome. In our empirical examples, the expected reward in an experiment is equal to the expected outcome or the expected success rate conditional on treatment assignment procedures. More complex reward functions could incorporate other costs (such as those associated with deploying the treatments) or benefits (such as those accruing to subjects as a result of the outcome).

Thompson Sampling

Thompson sampling is a heuristic approach to navigating the exploration–exploitation trade-off, facilitated by randomly assigning subjects to treatment arms according to their probability of returning the highest reward under a Bayesian posterior.[3] When there is not much information about which arm is best, the algorithm will facilitate exploration. As more information is gained, the best performing arms are increasingly exploited.

For ease of exposition we consider binary rewards, where each observation is either a success or a failure, $x \in \{0, 1\}$. Here, K arms have unknown success rates $\theta_1, \ldots, \theta_K$, following their respective Bernoulli distributions, with likelihoods

$$f_{X_1|\Theta_1}(x_1|\theta_1), \ldots, f_{X_K|\Theta_K}(x_K|\theta_K).$$

A researcher assigns some prior $f_{\Theta_k}(\theta_k)$ to the success rate of each arm. When researchers are initially

[2]For a history of the bandit problem and an overview of general approaches, see Berry and Fristedt (1985). For an updated overview of the general field of reinforcement learning, see Sutton and Barto (2018).

[3]We will focus primarily on Thompson sampling here, although there are many other algorithms, such as the upper confidence bound (UCB) algorithm, which selects the arm with the highest upper bound on an uncertainty interval around its estimated value, and the Epsilon-greedy algorithm, which selects the arm with the highest value most of the time, and assigns treatment randomly ε share of the time. The relative performance of each allocation rule depends on the time-horizon of the trial and the yardstick used to measure success: regret, statistical power, type I error rates (Villar, Bowden, and Wason 2015).

agnostic about the relative performance of the K arms, priors are distributed uniformly over the parameter space, which here is Beta(1, 1). If the researchers have prior beliefs or evidence regarding the performance of the arms, they may set the prior distributions accordingly. In each period t, treatment is assigned and observations are observed for each arm k, respectively. Let $n_{k,t}$ be the cumulative assignment to arm k and $X_k^{\{n_{k,t}\}} = (X_{[1]k}, \ldots, X_{[n_{k,t}]k})$ be the vector of responses under treatment arm k observed up until and including time t. The distribution of each Θ_k given the data $X_k^{\{n_{k,t}\}}$ in time t is then

$$f_{\Theta_k|X_k^{\{n_{k,t}\}}}\left(\theta_k|x_k^{\{n_{k,t}\}}\right) \propto f_{X_k^{\{n_{k,t}\}}|\Theta_k}\left(x_k^{\{n_{k,t}\}}|\theta_k\right) f_{\Theta_k}(\theta_k).$$

The Beta distribution is a conjugate prior for Binomial likelihoods, and consequently the posteriors follow a Beta distribution. The posterior α parameter is equal to one plus the total number of successes observed from that arm, and the posterior β parameter is equal to one plus the total number of failures observed from that arm.

In each period t, treatment is randomly assigned according to the probability of arms being best, that is,

$$P\left[\Theta_k = \max_k\{\Theta_1, \ldots, \Theta_K\}|\left(X_1^{\{n_{1,t}\}}, \ldots, X_K^{\{n_{K,t}\}}\right)\right],$$

and rewards are observed.[4] At the end of the period, the posterior is updated according to the successes and failures observed in that period, and the probability that each arm is best is recalculated. In the subsequent period, treatment assignment continues according to the updated probabilities.

Thompson sampling can be adapted to allow for drift in parameter values over time (Gupta, Granmo, and Agrawala 2011) or can account for reward probabilities that vary based on other variables that describe the context in which the action is taken (Agrawal and Goyal 2012). It can also be applied to more complex problems considered under reinforcement learning, where actions can affect future states, and information about rewards is delayed or sparse (Russo et al. 2017; Sutton and Barto 2018). In some applications, such as the ones considered here, adaptive trials end after a fixed period or when a predetermined number of subjects have participated in the trial. In other applications, the trial stops when any arm achieves a prespecified probability of being best; when used to establish statistical significance of effects, such stopping rules can run the risk of producing a false discovery, as the trial may stop if the best performing arm surpasses the target due to chance (Berman et al. 2018).

Here, our first objective is to select the best arm and estimate mean outcomes under this arm, not, as is common in the social sciences, to estimate average treatment effects, which is taken up in the next section. Indeed, if we assign treatment probabilities to all arms under Thompson sampling, and the control arm performs poorly relative to other arms, relatively few subjects will be assigned to the control arm, in which case estimates of the average treatment effect will typically have a larger standard error than under a static, balanced design (see simulations in SI D, p. 21).

Even if we are interested solely in estimating arm-specific means, Nie et al. (2017) demonstrate that sample means from adaptive experiments are prone to bias. (See SI C.2 p. 14, this bias is discussed in further detail in Villar, Bowden, and Wason 2015; Bowden and Trippa 2017.) For this reason, we use inverse probability weighting (IPW) estimators to account for bias introduced by sampling procedures. However, such estimators can exhibit large variance for arms with low sampling probabilities, which means that a standard static design may be preferable if precise evaluation of *all* treatment arms is the primary research objective.

An Algorithm for Adaptive Trials with a Control Condition

Researchers often seek to test whether one or more interventions outperform a control condition. Depending on the researchers' theoretical objectives, the control condition may involve a placebo, a business-as-usual treatment, or no intervention whatsoever. The inclusion of a control condition adds a layer of complexity to an adaptive trial.[5] When the aim is to gauge causal effects vis-à-vis a control group, the researcher must allocate sample to explore competing treatment arms while reserving sufficient sample for the control arm so that the resulting treatment-versus-control comparison is as precise as possible.

[4] We will use the term "probability of being best" to refer to the posterior probability that a given arm has the highest value of θ_k. For a worked numerical example, see SI C (p. 13). In practice, however, we generally estimate the value through simulation, taking a series of random draws from the posterior probability distributions of all arms, calculating the share of the series in which each arm had the highest draw, as implemented in the `bandit` package for `R` (Lotze and Loecher 2014).

[5] Methods for accounting for a control condition in adaptive trials have been considered before in clinical settings. Villar, Bowden, and Wason (2015) propose an approach in which an adaptive algorithm is used for treatment arms, but patients are assigned to a control condition with a fixed probability. Trippa et al. (2012) and Wason and Trippa (2014) also propose hybrid Bayesian randomization schemes, where assignment to the control condition depends on the cumulative sample assigned to a treatment arm and tuning parameters chosen by the researcher.

We propose a control-augmented adaptive algorithm to address this trade-off. In the first period, we assign treatment uniformly at random, as we would under Thompson sampling when we posit the same priors for all arms. The sample allocation probabilities for subsequent periods are then built from two component parts. The first part ensures that we sample sufficiently from the control condition. We calculate posterior probabilities of being best for each arm and identify the "current best arm" as the arm with the highest posterior probability. We then compare the cumulative sample assigned to the control condition to the cumulative sample assigned to the current best arm. If the cumulative sample assigned to the control condition is smaller, we calculate the portion of the subsequent batch that must be assigned to the control condition to achieve parity between the control sample and the sample of the current best arm. For example, if the current best arm is ahead of the control by three observations, and the expected batch size is 10, 30% of the treatment assignment probability will go to the control condition. We cap this probability at 90% to ensure that there is nonzero probability of assigning any of the treatment arms in each period. (This probability ceiling operates comparably to probability floors, as in, e.g., Dimakopoulou et al. 2017. The choice of 90% or some other high probability is arbitrary and simply guards against rare events; in practice, this constraint was seldom a binding constraint in our simulations with sufficiently large batches.)

The second part of the sampling probabilities navigates the exploration–exploitation trade-off across conditions. For the control, there is no trade-off, so we assign a fixed $1/K$ of the remaining probability to the control condition (where K is the number of treatment arms). The $(K-1)/K$ of the remaining probability is divided among treatment conditions proportional to their calculated posterior probabilities of being best, as in standard Thompson sampling.[6]

We assign treatments according to these probabilities in the next period, and the process of updating assignment probabilities begins again. A formalization of this algorithm is in SI A (p. 3), along with a worked example in SI C.3 (p. 18). In SI B.3 (p. 9), we discuss the theoretical properties of the algorithm.

[6]In SI A (p. 3), we discuss allowing control assignment to vary over successive batches by adjusting algorithm parameters. Here, we have emphasized allowing the number of subjects assigned to the control arm to 'catch up' to the number assigned to the best treatment arm to facilitate approximate balance, even as the arm we identify as 'best' changes across batches.

Hypothesis Testing in Adaptive Trials

Hypothesis testing under adaptive experimentation is complicated by the design feature that later treatment assignments and outcomes depend on earlier treatment assignments and outcomes. Hypothesis testing procedures such as *F*- or *t*-tests that compare average outcomes across arms typically assume that group mean estimates are statistically independent, but under adaptive designs they are not. The consequence of this dependence is that naive hypothesis tests will tend to be overconfident, yielding smaller *p*-values than is appropriate.

To conduct joint hypothesis tests, we propose a randomization inference procedure against the sharp null hypothesis of no differences across all arms for all units.[7] The test proceeds as follows. First, we obtain the *F*-statistic from an inverse-probability weighted regression of the outcome on indicators for all groups. Second, we simulate the distribution of the *F*-statistic under the sharp null hypothesis that each unit would express the same outcome as observed, regardless of treatment assignment. This null distribution is importantly different from the theoretical *F*-distribution implied by the nominal degrees of freedom: The simulated null distribution accounts for the dependence across units by allowing the randomization algorithm to adapt differently in each run of the simulation. Finally, we obtain a *p*-value by observing the proportion of simulated *F*-statistics under the sharp null that are more extreme than the observed *F*-statistic.[8][9]

The randomization inference approach will work for joint tests (such as the *F*-test) that compare across all K arms, but it will be inappropriate for pairwise comparisons across arms. The reason for this is that in a pairwise test, we want to test the null hypothesis of no

[7]For a textbook introduction to randomization inference, see Gerber and Green (2012, chapter 3) with a discussion of implementation under static designs. A superpopulation-based permutation test for adaptive trials similar in spirit to the one we discuss here is presented in Wei (1988).

[8]Although there is a relationship between the nulls under the joint test and the pairwise test of the best arm compared to control, we may fail to reject the null under either test and yet reject under the other. The added value of the joint test is that it serves as a robustness check and better accounts for the full realization of treatment assignment and response.

[9]Compared with adaptive designs, static designs are better powered to detect deviations from the null hypothesis of no effect for any unit in any arm, because under the null, static designs will estimate all group means with greater precision. In other words, if all arms are equally effective, the adaptive procedure will not find a best arm (because no arm is best), and it will add variance in the process.

TABLE 1 Iterated Simulation Statistics

Design			RMSE		Coverage	
Assignment algorithm	**Case**	**Best arm selected**	**Best arm**	**ATE**	**Best arm**	**ATE**
TS	1: Clear winner	0.968	0.021	–	0.958	–
	2: No clear winner	0.193	0.033	–	0.880	–
	3: Competing second best	0.715	0.025	–	0.956	–
Static	1: Clear winner	0.909	0.031	0.038	0.941	0.949
	2: No clear winner	0.180	0.024	0.033	0.935	0.947
	3: Competing second best	0.635	0.031	0.038	0.940	0.945
TS,	1: Clear winner	0.956	0.023	0.029	0.957	0.952
Control-Augmented	2: No clear winner	0.174	0.034	0.041	0.879	0.886
	3: Competing second best	0.683	0.029	0.035	0.946	0.937

Note: Assignment algorithms are Thompson sampling (TS), balanced static design (Static), and control-augmented Thompson sampling (TS, Control-Augmented). "Best arm selected" column presents the portion of simulations under which the true best arm was selected. RMSE is average root mean squared error of the estimate of the mean of the true best arm, and the average treatment effect of the true best arm relative to the control. Coverage is with respect to 95% confidence intervals around the estimate. In all cases one of the inferior arms with a true success rate of 0.10 is selected as the control comparison.

difference in outcomes across the two particular arms in question, but the hypothesis does not specify the outcome distributions under the remaining $K-2$ arms. This problem applies to static trials as well (Young 2019) but is even more vexing for adaptive trials because sampling probabilities depend on observed outcomes under all arms.

One approach is to restrict the use of standard t-tests to large studies. Under the control augmented design, when there is a unique best arm, asymptotically, standard t-tests will provide proper coverage (Zhang, Janson, and Murphy 2020). (Here and throughout, we use HC2 robust standard errors for confidence intervals, as implemented by the `estimatr` package, Blair et al. 2020.) The reason is that eventually the adaptive algorithm will converge to a pairwise comparison between the best arm and the control arm, although Zhang, Janson, and Murphy (2020) note that when the signal-to-noise ratio is small, undercoverage may still occur in finite samples. Our simulation evidence (presented in SI D p. 21) confirms that when there is a unique best arm and sufficient data, the confidence intervals for the difference-in-means test have correct coverage. When there is no best arm or when an adaptive study is too small to reliably detect a best arm, hypothesis testing becomes more complex and is an active area of statistical investigation. In SI B (p. 6), we provide an overview of this literature and proposed methods.

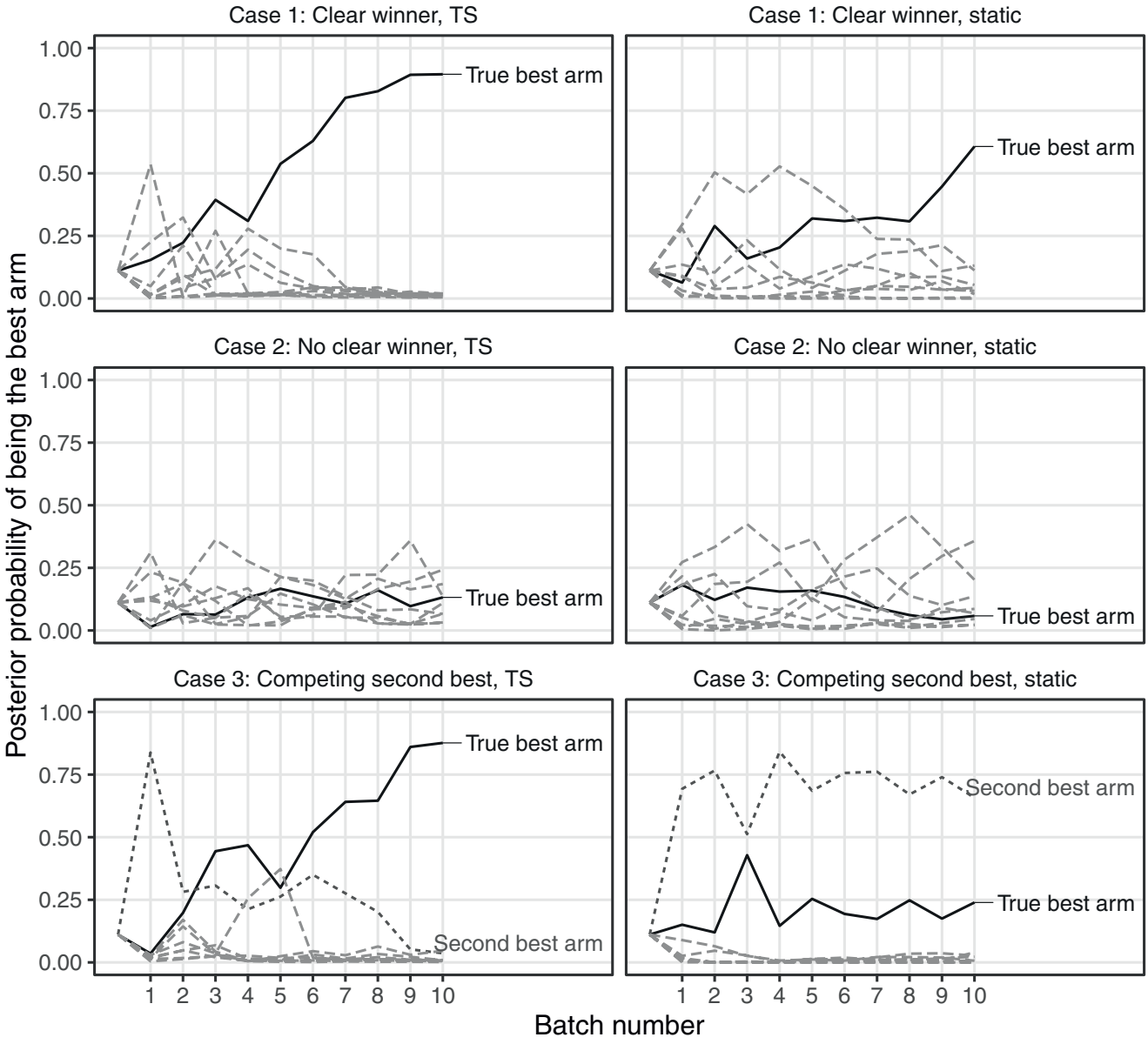
Simulations Illustrating How Adaptive Designs Work

We illustrate properties of adaptive designs under several hypothetical scenarios. In these scenarios, we simulate experiments sampling 100 observations for each of 10 periods, assigning treatment to nine arms according to the adaptive algorithms or a standard static design. In the adaptive algorithms, we set uniform priors for all arms, and arms are sampled with equal probability in the first period. The choice of 10 periods is arbitrary but anticipates the empirical examples presented below, which run for 10 days.

We illustrate development of a *single* simulated experiment in each scenario under Thompson sampling and a static design (Figure 1), and under the control-augmented design (Figure 2). We then present averages across 10,000 simulations (Table 1). In SI D, we repeat these simulations, varying the number of batches, the success rate of the best arm, and the size of the first batch.

We consider three scenarios. In the first case, there is a clear winner: One arm has a 0.20 success rate, and the remaining eight arms have success rates of 0.10. In the second case, there is no clear winner: The best arm has only a 0.11 success rate, and the remaining eight arms have success rates of 0.10. In the third case, the best arm is clearly superior to most other arms, but there is a

FIGURE 1 Simulated Posterior Probabilities over Time, Thompson Sampling and Static Designs



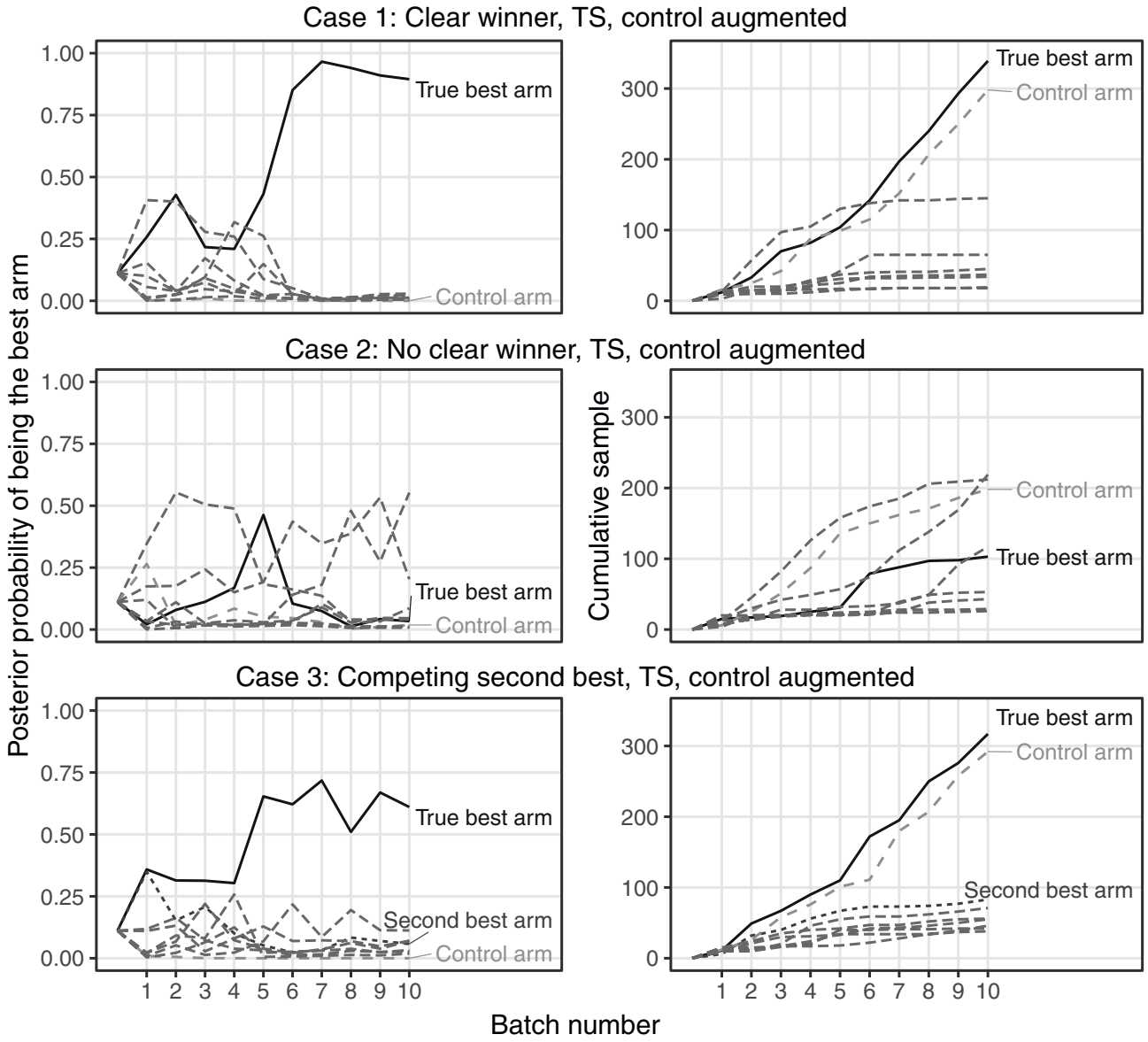
Note: Assignment algorithms are Thompson sampling (TS) and balanced static design (static). Success probabilities are as follows:
Case 1: (0.2, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1),
Case 2: (0.11, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1),
Case 3: (0.2, 0.18, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1).

competing second-best arm: The best arm has a 0.20 success rate, a second-best arm has a 0.18 success rate, and the remaining seven arms have success rates of 0.10.

In the single simulated experiment illustrating the first case (top panels of Figure 1), we see that the true best arm takes an early lead in the adaptive design. By the end of the 10-period experiment, the true best arm is assigned a 0.90 probability of being best in the adaptive design and a .61 probability of being best in the static trial. The "clear winner" scenario highlights the advantages of adaptive design over static design.

In the second case (center panels of Figure 1), the best arm is only very slightly superior to alternatives. In this "no clear winner" scenario, we do not correctly identify the true best arm in either the adaptive or static trials. Indeed, we assign the best arm only 0.13 probability of being best, whereas we assign an inferior arm 0.24 probability of being best. For the static experiment, we

FIGURE 2 Simulated Posterior Probabilities over Time and Cumulative Sample, Control-Augmented Adaptive Design



Note: The left panel represents posterior probability of being best, the right panel cumulative sample assigned to each arm. Success probabilities are as follows:
Case 1: (0.2, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1),
Case 2: (0.11, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1),
Case 3: (0.2, 0.18, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1).

assign the best arm 0.06 probability of being best, and an inferior arm a 0.36 probability of being best—this is a case where an inferior arm has overperformed relative to other arms merely by chance.

In the third case (bottom panels of Figure 1), the adaptive design assigns the best arm a 0.88 probability of being best and the second best arm a 0.04 probability of being best. The static design accords the second-best arm a 0.66 probability of being best, but gives the true best arm only a 0.24 probability of being best. This "competing second-best" scenario illustrates where adaptive designs may help us differentiate between similarly successful treatments.

We repeat these simulations under our proposed control-augmented adaptive algorithm, presented in Figure 2, where in each case one of the arms with a

true success rate of 0.10 is designated as the control comparison. Posterior probabilities under the control-augmented design are similar to those under standard batch-wise Thompson sampling presented in Figure 2. However, whereas under standard Thompson sampling treatment assignment probabilities would correspond to the posterior probabilities, under the control-augmented algorithm, the cumulative sample assigned to the control condition is nearly matched to the presumed best arm. In the first and third cases, this presumed best arm is indeed the true best arm in our illustrated simulation. In the no clear winner case, we observe a similar pitfall as under standard Thompson sampling above, and have mistakenly assigned the highest posterior probability of being best to an inferior arm.

A key feature of adaptive designs is that the probability of assignment to each condition is dependent on observed outcomes and varies over time. To account for bias due to the dependent nature of the assignment probabilities, throughout we use IPW estimators, under which each observation is weighted by the inverse of the probability of assignment to the condition that it is in (see Gerber and Green 2012 for an introduction to IPW; see Bowden and Trippa 2017 for an investigation of IPW estimators in this setting).

These simulations show that adaptive designs tend to outperform static designs in settings where the research goal is to find the best performing arm. Under favorable conditions (the first and third cases), the adaptive designs do better than the static design, and under unfavorable conditions (the second case), the standard Thompson sampling design does slightly better and the control-augmented version does only slightly worse. However, this trend does not necessarily follow to other research goals. In estimation, there are potentially large gains in precision from adaptivity for the best performing arms, due to the alignment of sampling probability with arm performance. We see this reflected in the RMSE of the best arm in the first and third cases. However, as we use IPW to facilitate unbiased estimation, fluctuations in sampling probabilities may increase the variance of our estimates: Even if we assign more total observations to a given arm under an adaptive setting than under a static design, some periods with lower sampling probabilities can result in an overall higher variance of the estimate. We see this occur in the second case, where although we assign on average a larger portion of the sample to the true best arm in the adaptive setting as compared to the static design, our estimates of outcomes under the best arm are less precise. Considering the estimate of the average treatment effect of the best arm relative to the control, the control-augmented design represents a large improvement over the static design in the first and third cases. However, if researchers are genuinely interested in all arms equally, then a static trial that assigns the same fraction of subjects to each arm will clearly dominate an adaptive trial.

Empirical Application: Finding the Best Performing Treatment Arm

Simulations provide useful intuition about the conditions under which adaptive designs are helpful, but empirical applications allow us to illustrate how these designs may be implemented and analyzed in settings relevant to political scientists. Our first empirical applications address the wording of ballot measures. For this study, we recruited 1,000 subjects from the Amazon Mechanical Turk (MTurk) marketplace.[10]

Design

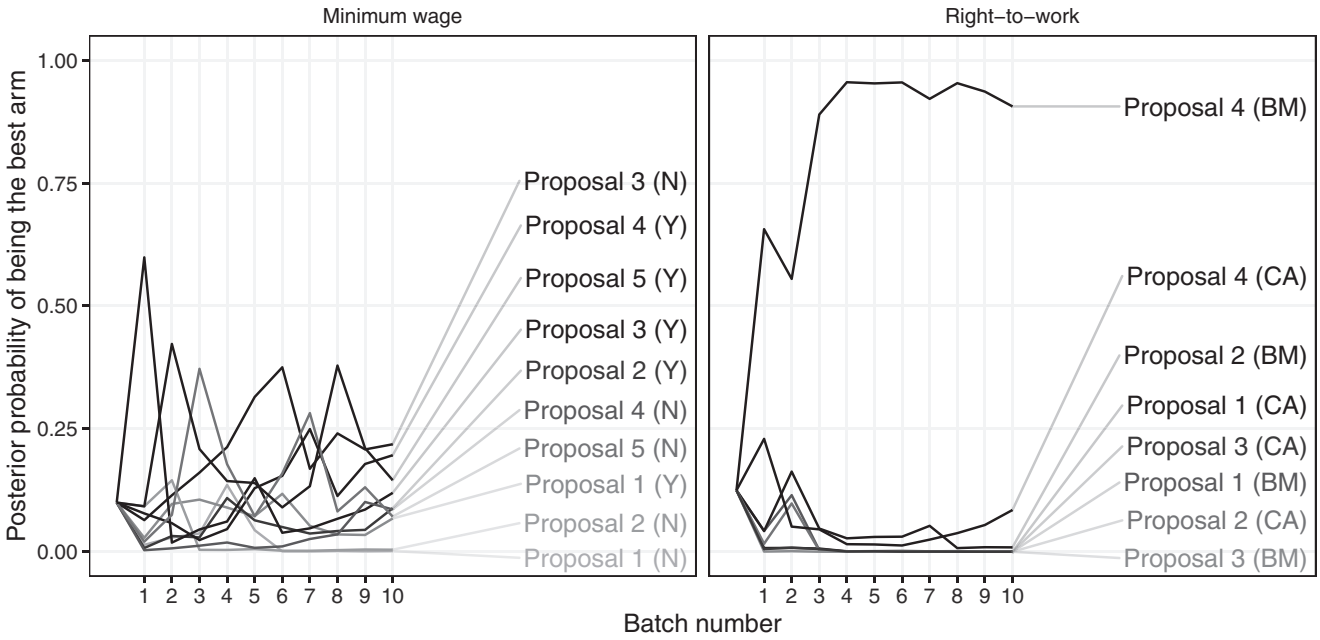
After answering a series of demographic questions, all subjects rated two ballot measures, one on minimum wage and one on right-to-work laws. We adapted the wording of the proposed measures from real proposals, making only minor changes to facilitate consistency across arms. Our objective in this setting is to estimate mean outcomes under the most popular ballot measure; there is no comparison to a control condition, and so we use standard Thompson sampling. In order to facilitate comparisons between alternative sampling methods, we implemented a composite adaptive-static design: For each type of ballot measure, subjects were assigned treatment according to Thompson sampling with 90% probability, and according to balanced simple random assignment with 10% probability.

The minimum wage treatments were drawn from ballot measures proposed in Colorado, Florida, Illinois, Nevada, and New Jersey. We generated two versions of each of these five proposals, varying whether the current value of the minimum wage was displayed, resulting in 10 unique minimum wage treatments.[11] The

[10] Convenience samples obtained on MTurk are far from representative of the national population but do provide a fertile testing ground for experimental studies. Recent meta-analyses have revealed a close correspondence of experimental estimates obtained on MTurk and probability samples (Mullinix et al. 2015; Coppock 2018; Coppock, Leeper, and Mullinix 2018).

[11] Minimum wage rates are presented in SI E (p. 33). For states that do not have a state minimum wage, we imputed the federal minimum wage value.

FIGURE 3 Study 1, Over Time Posterior Probabilities



Note: "Y" versions of the minimum wage proposals include the current minimum wage and "N" versions do not. 'CA' versions of the right-to-work proposals are described as 'constitutional amendments' and 'BM' ('ballot measure') versions are not.

right-to-work treatments were adapted from ballot measures in Missouri, North Dakota, Oklahoma, and South Dakota. For each of these, we created versions that did or did not describe the ballot measure as a "constitutional amendment," resulting in eight unique right-to-work treatments. For both rating tasks, the outcome question asked was, "If this measure were on the ballot in your state, would you vote in favor or against?" The outcome is defined as a success if the subject responds that they would vote in favor of the measure. The full text of all treatments is presented in Table 2. Here, the "best" treatment is the one with the highest associated success rate; there is no comparison to a baseline control condition. Our study ran from June 21 to June 30, 2018, and we paid subjects $1 each for their participation.

Results

We present two sets of results. The first, in Figure 3, is the over time development of the posterior probability that each arm is best. The second, in Figure 4, is a straightforward comparison of the inverse-probability weighted average approval of each proposal.

The minimum wage study yielded no clear winner. The winning arm, by a hair, was Proposal 3 (B, without current minimum wage), with a probability of being best of 0.219 and an estimated mean of 0.895 over 183 respondents. This arm was closely followed by Proposal 5 (Y and N, where "Y" is with and "N" is without minimum wage) and Proposal 4 (Y, with current minimum wage). Of 10 arms, only two had success rates under 0.8; with similarly high probabilities of success across several arms, the best arm was not easily distinguishable. The randomization inference *F*-test yielded a *p*-value of 0.429, indicating that we cannot reject the sharp null that support is unaffected by wording nuances.

By contrast, divergence in mean outcomes is readily apparent in the right-to-work experiment, which early on revealed a standout arm. Proposal 4 (framed as a ballot measure) ended with a 0.906 probability of being best. The second-best arm was also Proposal 4 (framed as a constitutional amendment) with a probability of being best of 0.085. The posterior probabilities are based on the unweighted number of successes and number of trials, but the probability weighted estimates are 0.926 and 0.934 over 721 and 82 respondents, respectively, and the difference in the two estimates is statistically indistinguishable from zero. Both versions of Proposal 4 are highly successful, and we note that the content of this measure, and its initial emphasis on protection of the rights to "life, liberty or property," may make it more appealing than the alternatives that address union membership first. The randomization inference *F*-test yielded

TABLE 2 Study 1, Treatments and Outcome Measures

	Minimum wage	Right to work
Question text	Imagine that the following ballot measure were up for a vote in your state. [**ballot measure text**]. If this measure were on the ballot in your state, would you vote in favor or against? [I would vote in favor of this measure; I would vote against this measure]	Imagine that the following ballot measure were up for a vote in your state. [**ballot measure text**]. If this measure were on the ballot in your state, would you vote in favor or against? [I would vote in favor of this measure; I would vote against this measure]
Proposal 1	The measure would: increase the minimum wage [**from {current}**] to {current + 1} per hour, adjusted annually for inflation, and provide that no more than $3.02 per hour in tip income may be used to offset the minimum wage of employees who regularly receive tips.	The measure would [**amend the State Constitution to**]: prohibit, as a condition of employment, forced membership in a labor organization (union) or forced payments of dues or fees, in full or prorata ("fair-share"), to a union. The measure will also make any activity that violates employees' rights provided by the bill illegal and ineffective and allow legal remedies for anyone injured as a result of another person violating or threatening to violate those employees' rights. The measure will not apply to union agreements entered into before the effective date of the measure, unless those agreements are amended or renewed after the effective date of the measure.
Proposal 2	The measure would: raise the minimum wage [**from {current}**] to {current + 1} per hour effective September 30, 2021. Each September 30 thereafter, minimum wage shall increase by $1.00 per hour until the minimum wage reaches {current + 5} per hour on September 30, 2026. From that point forward, future minimum wage increases shall revert to being adjusted annually for inflation starting September 30, 2027.	The measure [**reads/would amend the State Constitution to read**]: The right of persons to work may not be denied or abridged on account of membership or nonmembership in any labor union or labor organization, and all contracts in negation or abrogation of such rights are hereby declared to be invalid, void, and unenforceable.
Proposal 3	The measure reads: Shall the minimum wage for adults over the age of 18 be raised [**from {current}**] to {current + 1} per hour by January 1, 2019?	The measure would [**amend the State Constitution to**]: ban any new employment contract that requires employee to resign from or belong to a union, pay union dues, or make other payment to a union. Required contributions to charity or other third party instead of payments to union are also banned. Employees must authorize payroll deduction to unions. Violations of the section is a misdemeanor.
Proposal 4	The measure would: raise the minimum wage [**from {current}**] to {current + 1} per hour worked if the employer provides health benefits, or {current + 2} per hour worked if the employer does not provide health benefits.	The measure [**reads/would amend the State Constitution to read**]: No person shall be deprived of life, liberty, or property without due process of law. The right of persons to work shall not be denied or abridged on account of membership or nonmembership in any labor union, or labor organization.
Proposal 5	The measure would: raise the State minimum wage rate [**from {current}**] to at least {current + 1} per hour, and require annual increases in that rate if there are annual increases in the cost of living.	

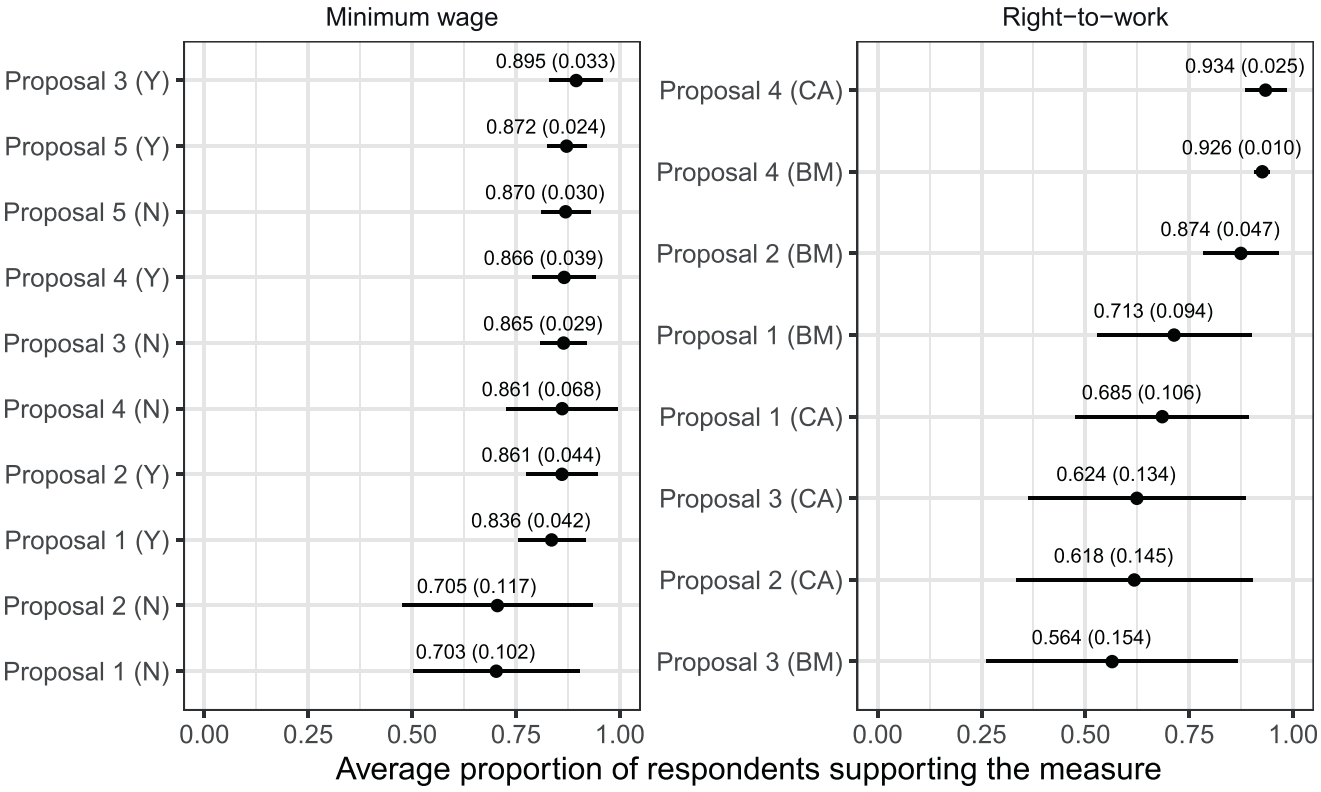
Note: Boldface text indicates randomly varied elements.

a p-value of 0.199, indicating that we again cannot reject the sharp null that support is unaffected by ballot wording. The failure to reject the null in spite of the large range in arm means is some indication that the randomization inference test is relatively underpowered in cases where an adaptive trial allocates the majority of respondents to a single treatment arm.

We can use these results to inform guesses about how our experiment would have fared if we had used a standard static design instead of the adaptive design. The static design would have sampled each of the 10 arms in the minimum wage experiment in expectation 100 times each and each of the eight arms in the right-to-work experiment in expectation 125 times each. Treating observed success rates as the truth in simulations of the minimum wage experiment, we picked the best proposal 47% of the time in adaptive experiments and 37% of the time in static experiments. In the right-to-work experiment, where there are two strong contenders, we picked the best proposal 65% of the time in adaptive experiments and 58% of the time in static experiments.

Typically when analyzing an experiment, we estimate arm-specific means and provide associated standard errors. Considering Figure 4, we note that a feature of the adaptive design is that the proposals with the highest success rates also have the smallest standard errors, as these arms tend to receive more subjects than arms with lower success rates. This feature may be desirable when estimation of average performance of the best arm is considered a priority by the researcher and estimates of average success rates of poorly performing arms are not of particular interest. For the minimum wage experiment, standard errors around our estimate of the success rate for the best arm in a simulated static design would have been, on average, 90% as large as those under a comparable adaptive design. This reflects the cost in variance of due to fluctuations in inverse probability weights in the adaptive design, when the best arm does not quickly achieve a high sampling probability. For the right-to-work experiment, standard errors on the success rate of the best arm in a static experiment would have been on average 133% as large as those under a comparable adaptive design. When there is a standout arm, the adaptive design appears to

FIGURE 4 Study 1, Mean Vote Outcomes



Note: Estimates are inverse probability weighted. Standard errors are heteroskedasticity consistent HC2 corrected. "Y" versions of the minimum wage proposals include the current minimum wage and "N" versions do not. "CA" versions of the right-to-work proposals are described as constitutional amendments and "BM" versions are described as ballot measures.

offer advantages in terms of the precision with which the best performing arm's success rate is evaluated.

Empirical Application: Adaptive Trials with a Control Condition

This section illustrates our algorithm for adaptive trials with a control condition using an example from the growing literature on partisan bias in public opinion. For decades, scholars have noted the "perceptual screen" through which party loyalists absorb and retain factual information (Campbell et al. 1960, p. 133). Partisans are predictably buoyant or gloomy depending on which party holds the presidency. For example, Democrats and Republicans have markedly different recollections of past economic performance (Bartels 2002) and assessments of current and future conditions (Gerber and Huber 2010). A lively theoretical and empirical literature focuses on the question of what can be done to reduce partisan bias. To what extent do survey respondents rein in their partisan impulses when offered financial incentives for accuracy (Bullock et al. 2015; Khanna and Sood 2018; Prior et al. 2015)? What if respondents are urged to put aside partisan biases when answering factual questions (Prior et al. 2015)? What if they are given extra time to reflect on their answers or to consult factual sources (Prior and Lupia 2008)? All of these methods for inducing factual accuracy have received support, and the purpose of the experiment that follows is to discern which works best and whether any of these methods produce appreciable gains vis-à-vis an untreated control group.

Design

Our outcome measures are responses to three factual questions listed in Table 3. These questions address trends in budget deficits, black unemployment, and farm income between the Obama and Trump years. As expected, responses to these questions are highly correlated with respondent party. For example, restricting our attention to those asked the control version of each question, we see that 24% of Republicans and 62% of Democrats reported that the deficit had grown under Trump; 75% and 34% reported that black unemployment had decreased; and 22% and 59% said that farm income had declined. Our aim, however, is not to compare Republicans to Democrats. Rather, we conduct separate randomized trials within each partisan subgroup to see what encourages increased accuracy.

Drawing on the recent literature, we implement a six-arm trial, with a control condition and each of five treatment arms representing a somewhat different theoretical approach. As shown in Table 4, the treatment conditions amount to variations in the instructions and encouragements given to respondents. The Lottery treatment is a financial incentive, whereby accurate answers increase respondents' odds of winning a $100 gift certificate. Second, the Accuracy treatment instructs respondents to "answer these questions as accurately as you can." The Directional treatment encourages respondents to put aside their partisan bias: "We know from past surveys that people tend to root for their own political party, but often the right answer is not the one that favors one's own party. Please be sure to think objectively – without any partisan bias – about true economic conditions before answering." The Extra Time condition allots respondents 45 seconds to formulate an answer (and prevents them from advancing to the next question during that time). Finally, the Google treatment encourages respondents to look up the correct answers: "We know from past surveys that sometimes people use Google to look up the answers to questions like these. For this survey, it is OK to use Google!" Respondents were assigned the same treatment condition for all three questions, and our analysis therefore must account for the fact that the assignment for each question–respondent pair is effectively clustered by respondent. Our outcome was scored 1 if the respondent supplied the correct answer and 0 otherwise. To avoid posttreatment bias, we scored respondents who did not provide any answer at all as 0 as well (Coppock 2019; Montgomery, Nyhan, and Torres 2018).

Using the Lucid platform, we gathered approximately 300 observations per day for a total of 10 days and updated treatment assignment probabilities after each batch of data collected.[12] Our study ran from November 27 to December 10, 2019, and subjects were paid $1 for their participation. In the first batch, we assigned all treatment conditions and the control with equal probability. In subsequent periods, allocation to experimental group was conducted separately for Democrats and Republicans, whose party identification is measured using the conventional ANES wording early in the survey, prior

[12] Like MTurk samples, Lucid samples are online convenience samples that are not necessarily representative of the national population. That said, respondents on Lucid are quota sampled to match U.S. census demographic margins, ensuring sample diversity on many important dimensions (Coppock and McClellan 2019).

TABLE 3 Study 2, Questions and Outcome Measures

	Deficit	Net farm income	Black unemployment
Question text	In 2016, the last year of the Obama presidency, the deficit was 3.1% of GDP. We are now 3 years into the Trump presidency. Would you say that the deficit has gotten better, gotten worse, or stayed the same in relation to the GDP?	The U.S. Department of Agriculture tracks the financial well-being of American farmers. An important indicator is Net Farm Income, which measures how much money farmers make after expenses. In 2013, 4 years into the Obama presidency, Net Farm Income was 120 billion. We are now 3 years into the Trump presidency. Would you say that Net Farm Income has gotten worse, gotten better, or stayed the same?	The U.S. Bureau of Labor Statistics counts a person as unemployed if they are not currently working and are looking for work. At the end of 2011, 3 years into the Obama presidency, the unemployment rate for black Americans was 15%. We are now 3 years into the Trump presidency. Would you say that the unemployment rate for black Americans has gotten worse, gotten better, or stayed the same?
Binary response	[Gotten worse, Gotten better, Stayed the same] Coded as 1 if correct, 0 otherwise		
Continuous response	If you had to guess, what is the current deficit as a percentage of GDP? [numeric entry]	If you had to guess, what is the value of Net Farm Income this year, in billions? [numeric entry]	If you had to guess, what is the current unemployment rate for black Americans? [numeric entry]
Answers	In fiscal year 2019, the federal government estimates that the deficit will be 5.1% of GDP, which means that the deficit has **gotten worse** since 2016.	In fiscal year 2019, Net Farm Income was 90 billion, which means that Net Farm Income has **gotten worse** since 2013.	As of September 2019, the unemployment rate for black Americans was 5.5%, which means that the unemployment rate for black Americans has **gotten better** since 2011.

to random assignment.[13,14] Thus, we have in effect two separate adaptive trials, one for each partisan group.

[13]We could not know exactly how many units would enter in each batch, because we split fixed-size batches based on self-reported party identification that is revealed during the course of the survey. To account for this complication, we based allocation procedures on the expected batch size.

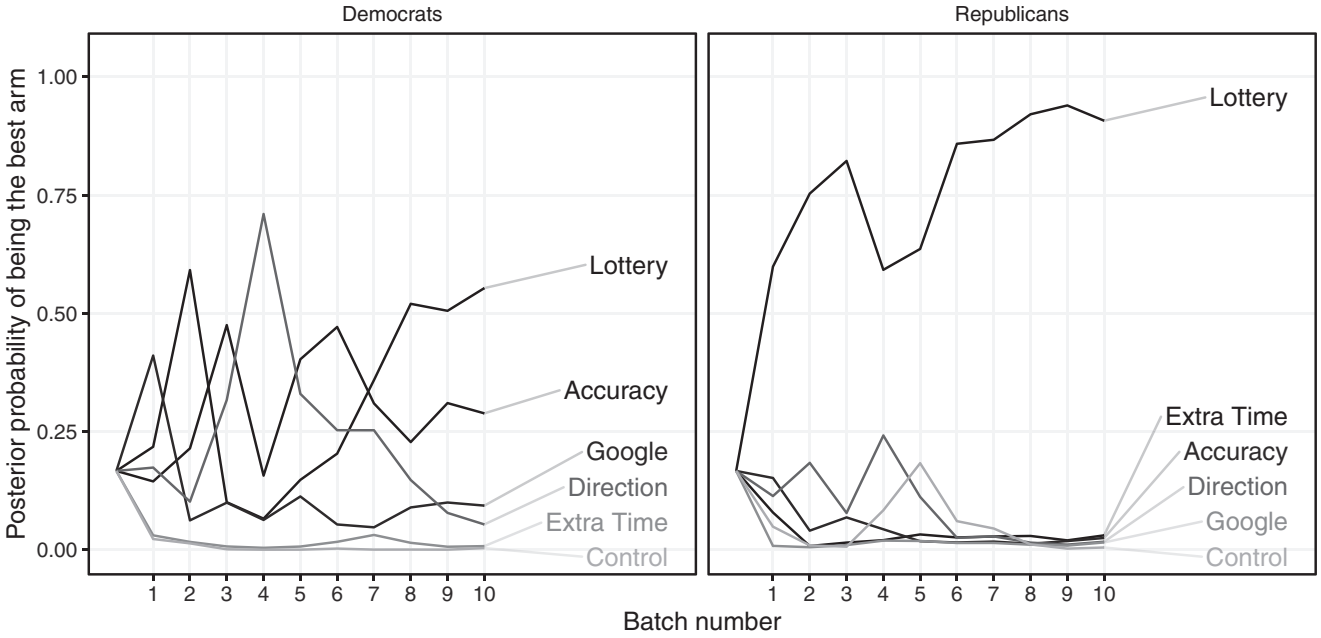
[14]We coded independent "leaners" as partisans. "Pure" independents were allocated separately as part of their own adaptive trial, but their numbers are small ($n = 568$), so we report these and other additional analyses in SI F (p. 34).

Turning first to the Republican respondents, we see from Figure 5 that the lottery incentive led the entire way and was the only treatment arm that outperformed the control. In the end, the unadjusted estimate indicates a 4.1 percentage point increase in accuracy with a 1.7 percentage point standard error, as shown in Figure 6. The fact that we obtain a standard error as small as 1.7 reflects the fact that the adaptive design allocated 552 of 1,258 Republicans to the Lottery condition, with another 493 assigned to control. Had we implemented a static design with equal allocation to all six conditions, based on

TABLE 4 Study 2, Treatments and Reminders

	Subjects assigned to the control condition proceed directly to the outcome questions. Subjects assigned to one of the five treatment conditions first see a preamble that delivers the treatment. Then, when they proceed to the outcome questions, they are reminded of the treatment.	
	Preamble	**Reminder**
Lottery	In this next section, you will be asked some questions about how well the economy is doing. As you probably know, the government gathers a lot of statistical information about the economy. Not everyone pays attention to these statistics or remembers them after they have heard them. In this study we would like to learn whether this information is finding its way to the general public. For each correct answer, we will award you an extra chance to win a lottery for a $100 Amazon gift card. You will earn an additional extra chance for every correct answer you give. The more questions you answer correctly, the better your chances are to win. At the end of this study, you will see a summary of how many questions you answered correctly.	As a reminder, for each correct answer, we will award you an extra chance to win a lottery for a $100 Amazon gift card. You will earn an additional extra chance for every correct answer you give. The more questions you answer correctly, the better your chances are to win. At the end of this study, you will see a summary of how many questions you answered correctly.
Accuracy	In this next section, you will be asked three questions about how well the economy is doing. As you probably know, the government gathers a lot of statistical information about the economy. Not everyone pays attention to these statistics or remembers them after they have heard them. In this study we would like to learn whether this information is finding its way to the general public. The questions that follow have right and wrong answers. In order for your answers to be most helpful to us, it is really important that you answer these questions as accurately as you can. At the end of this study, you will see a summary of how many questions you answered correctly.	As a reminder, these questions have right and wrong answers. In order for your answers to be most helpful to us, it is really important that you answer these questions as accurately as you can. At the end of this study, you will see a summary of how many questions you answered correctly.
Direction	In this next section, you will be asked some questions about how well the economy is doing. As you probably know, the government gathers a lot of statistical information about the economy. Not everyone pays attention to these statistics or remembers them after they have heard them. In this study we would like to learn whether this information is finding its way to the general public. We know from past surveys that people tend to root for their own political party, but often the right answer is not the one that favors one's own party. Please be sure to think objectively—without any partisan bias – about true economic conditions before answering.	As a reminder, we know from past surveys that people tend to root for their own political party, but often the right answer is not the one that favors one's own party. Please be sure to think objectively—without any partisan bias—about true economic conditions before answering.
Extra time	In this next section, you will be asked some questions about how well the economy is doing. As you probably know, the government gathers a lot of statistical information about the economy. Not everyone pays attention to these statistics or remembers them after they have heard them. In this study we would like to learn whether this information is finding its way to the general public. So that you are not rushed in any way, we will give you extra time to complete each set questions. The "next" button will appear after 45 seconds so that you will have plenty of time to consider your answers.	As a reminder, the "next" button will appear after 45 seconds so that you will have plenty of time to consider your answers
Google	In this next section, you will be asked some questions about how well the economy is doing. As you probably know, the government gathers a lot of statistical information about the economy. Not everyone pays attention to these statistics or remembers them after they have heard them. In this study we would like to learn whether this information is finding its way to the general public. We know from past surveys that sometimes people use Google to look up the answers to questions like these. For this survey, it is OK to use Google! We just want to know what site you went to, so we'll ask you to copy-paste the link in. Of course, there is no need to search for the answer if you do not want to.	As a reminder, we know from past surveys that sometimes people use Google to look up the answers to questions like these. For this survey, it is OK to use Google! We just want to know what site you went to, so we'll ask you to copy-paste the link in. Of course, there is no need to search for the answer if you do not want to.

FIGURE 5 Study 2, Overtime Posterior Probabilities



Note: Respondents are coded as Democrat if their response is 1, 2, or 3 on the 7-point partisanship scale, and Republican if their response is 5, 6, or 7. Posterior probabilities are updated after each day's data collection according to the control-augmented Thompson sampling algorithm.

simulations the standard error would have been on average 2.5 percentage points (we present estimates and confidence intervals from simulated experiments under the static design in SI F.3, p. 36). One of the advantages of a control-augmented adaptive design in this context is that one emerges with a more precise assessment of the leading arm's causal effect. In this particular case, the adaptive design with the smaller standard error allows us to declare the 4.1-point average effect of the accuracy treatment statistically significant ($p = 0.014$), whereas under that static design the effect would not have attained statistical significance ($p = 0.152$). Using randomization inference, we reject the null of no difference across treatments ($p = 0.028$).

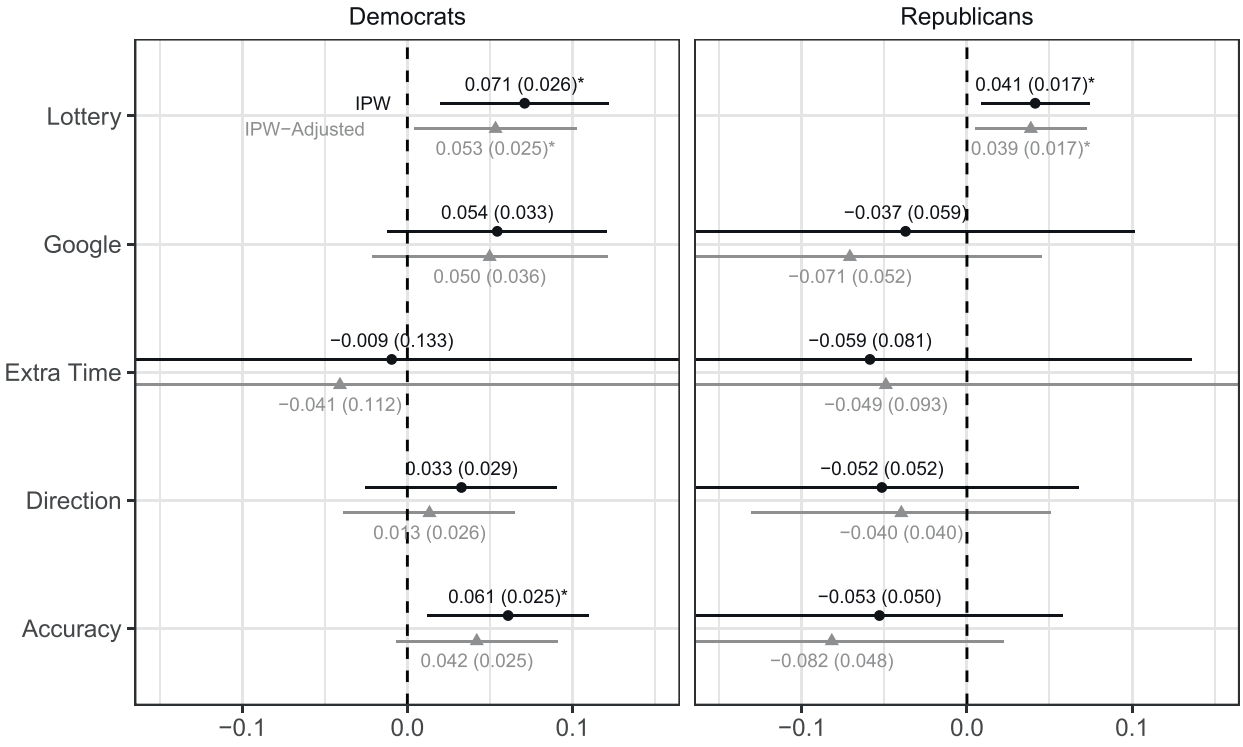
The trial involving Democrats illustrates how the gains from adaptive design may be more muted when multiple treatment arms generate similarly strong effects. Although Lottery won with a narrow lead, Democrats also responded to the Accuracy and Google encouragements. Thus, the subject pool was spread relatively evenly across conditions, with Extra Time being the sole arm that was ruled out early on. In the end, Lottery was significantly better than the control by 7.1 percentage points (unadjusted, $p = 0.007$) or 5.3 percentage points (adjusted, $p = 0.03$). Using randomization inference, we again reject the null ($p = 0.013$).[15]

Discussion

The growth and development of experimentation in the social sciences has led to increasing sophistication in the design of multiarm trials. Although the adaptive allocation of subjects to treatment arms over time adds complexity to a trial's implementation and analysis, the payoff may be considerable. When one arm is truly superior to the others, an adaptive trial can locate the winning arm more reliably than a static design. Moreover, because the adaptive trial allocates more sample to the winning arm, the experimenter learns more about the attributes of the winner at the conclusion of the study. Our simulations and the empirical example of right-to-work ballot

[15]Had we used the batched OLS approach proposed by Zhang, Janson, and Murphy (2020) discussed in SI B.1 (p. 6) to account for nonnormality, the *p*-values for the ATE of the Lottery condition rise slightly, with $p = 0.026$ for Republicans and $p = 0.018$ for Democrats.

FIGURE 6 Study 2, Average Treatment Effect Estimates



Note: Estimates are inverse probability weighted. Standard errors are bias-reduced linearization CR2 adjusted, clustered at the respondent level. Covariate-adjusted estimates control for age, gender, education, race, political ideology, political attentiveness, and partisanship.
*$p < 0.05$.

measures illustrate just how valuable adaptive designs can be in the context of a truly superior arm. The level of public support for the winning right-to-work ballot measure was estimated with a standard error that was 75% as large as would have been the case under a static design.

The adaptive allocation of subjects, however, is of less value when no treatment arm truly stands above the others. In such cases, adaptive allocation follows clues that are the product of sampling variability rather than the true superiority of an arm. As the minimum wage application suggests, at best an adaptive design winnows out some inferior arms. In this application, however, the cost in variance from extreme inverse probability weights resulted in a less precise estimate of the best arm outcome.

This point also holds for studies in which the aim is to compare treatment arms to an untreated control group. One component of the design aims to locate the best performing treatment arm, but the offsetting component ensures that the control group always receives ample subjects regardless of how it performs over the trial. When one treatment arm is truly superior, this design will allocate substantially more subjects to it and will therefore render a more precise estimate of the treatment effect vis-à-vis the control group. On the other hand, the gains may be negligible if the treatment arms are in fact similarly effective.

One important research frontier is the efficient allocation of sample in the context of highly factorial designs like conjoint experiments. Because the number of possible treatment arms is large relative to the number of subjects, adaptive design alone may be unable to isolate the best treatment combination with high probability over a fixed data collection schedule. In this case, adaptive design requires the assistance of modeling assumptions, such as additive effects, to reduce the set of promising treatment combinations. In the SI, we present an empirical demonstration of this approach.

Another important research frontier is designing adaptive trials when the true underlying performance of the treatment arms is believed to be changing over time. For example, an adaptive trial designed to gauge the political campaign ad that attracts the most support for the advertising candidate might operate in a fast-moving environment in which a candidate's popularity erodes over time. An adaptive trial might start out with one front-running ad, allocate more sample to it, and then see

the apparent performance of that ad deteriorate as the candidate's popularity drifts downward. Future research, building on the work of Granmo and Berg (2010) and Gupta, Granmo, and Agrawala (2011), will need to grapple with the added complexity of dynamics in both the design and analysis of adaptive trials.

A final challenge is to expand the framework of response-adaptive algorithms to potentially change the portfolio of treatment arms. In principle, an algorithm could inform researchers that it is time to develop new treatment arms because the existing ones are inadequate, either because all treatment arms perform below some preset standard or because none performs better than the control arm. This approach is akin to what researchers currently do informally when they conduct a pilot study and conclude that the results do not look sufficiently promising to warrant further experimentation with the treatments at hand. Under an expanded adaptive framework, the scope of exploration widens. Arms are compared not only to one another but also to the potential value of arms that have yet to be tested.

References

Agrawal, Shipra, and Navin Goyal. 2012. "Thompson Sampling for Contextual Bandits with Linear Payoffs." arXiv e-prints arXiv:1209.3352v4 [cs.LG].

Bartels, Larry. 2002. "Beyond the Running Tally: Partisan Bias in Political Perceptions." *Political Behavior* 24(2): 117–50.

Berman, Ron, Leonid Pekelis, Aisling Scott, and Christophe Van den Bulte. 2018. "p-Hacking and False Discovery in A/B Testing." SSRN, accessed October 4, 2020, https://doi.org/10.2139/ssrn.3204791.

Berry, Donald A., and Bert Fristedt. 1985. *Bandit Problems: Sequential Allocation of Experiments*. New York and London: Chapman and Hall.

Blair, Graeme, Jasper Cooper, Alexander Coppock, Macartan Humphreys, and Luke Sonnet. 2020. *estimatr: Fast Estimators for Design-Based Inference*. R package version 0.22.0. https://CRAN.R-project.org/package=estimatr

Bowden, Jack, and Lorenzo Trippa. 2017. "Unbiased Estimation for Response Adaptive Clinical Trials." *Statistical Methods in Medical Research* 26(5): 2376–88.

Bullock, John G., Alan S. Gerber, Seth J. Hill, Gregory A. Huber, et al. 2015. "Partisan Bias in Factual Beliefs about Politics." *Quarterly Journal of Political Science* 10(4): 519–78.

Campbell, Angus, Philip E. Converse, Warren E. Miller, and Donald E. Stokes. 1960. *The American Voter*. New York: Wiley.

Chin, Richard. 2016. *Adaptive and Flexible Clinical Trials*. Boca Raton: CRC Press.

Chow, Shein-Chung, and Mark Chang. 2008. "Adaptive Design Methods in Clinical Trials—A Review." *Orphanet Journal of Rare Diseases* 3(1): 11.

Coppock, Alexander. 2018. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* 1–16.

Coppock, Alexander. 2019. "Avoiding Post-Treatment Bias in Audit Experiments." *Journal of Experimental Political Science* 6(1): 1–4.

Coppock, Alexander, and Oliver A. McClellan. 2019. "Validating the Demographic, Political, Psychological, and Experimental Results Obtained from a New Source of Online Survey Respondents." *Research & Politics* 6(1): 2053168018822174.

Coppock, Alexander, Thomas J. Leeper, and Kevin J. Mullinix. 2018. "Generalizability of Heterogeneous Treatment Effect Estimates Across Samples." *Proceedings of the National Academy of Sciences* 115(49): 12441–46.

Dimakopoulou, Maria, Zhengyuan Zhou, Susan Athey, and Guido Imbens. 2017. "Estimation Considerations in Contextual Bandits." arXiv e-prints arXiv:1711.07077v4 [stat.ML].

Dimmery, Drew, Eytan Bakshy, and Jasjeet Sekhon. 2019. "Shrinkage Estimators in Online Experiments." arXiv e-preprints arXiv:1904.12918v1 [stat.ME].

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton.

Gerber, Alan S., and Gregory A. Huber. 2010. "Partisanship, Political Control, and Economic Assessments." *American Journal of Political Science* 54(1): 153–73.

Graepel, Thore, Joaquin Quiñonero Candela, Thomas Borchert, and Ralf Herbrich. 2010. "Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine." In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010) (ICML-10)*, June 21–24, 2010, Haifa, Israel, eds. J. Fürnkranz and T. Joachims. Omnipress, pp. 13–20. Accessed October 4, 2020, https://discovery.ucl.ac.uk/id/eprint/1395202

Granmo, Ole-Christoffer, and Stian Berg. 2010. "Solving Non-Stationary Bandit Problems by Random Sampling from Sibling Kalman Filters." In *23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2010, Cordoba, Spain, June 1-4, 2010, Proceedings, Part III (IEA-AIE 2010)*, eds. Nicolás García-Pedrajas, Francisco Herrera, Colin Fyfe, José Manuel Benítez, and Moonis Ali. Vol. 6098 of *Lecture Notes in Computer Science* Berlin, Heidelberg: Springer-Verlag, pp. 199–208. Accessed October 4, https://doi.org/10.1007/978-3-642-13033-5_21.

Gupta, Neha, Ole-Christoffer Granmo, and Ashok Agrawala. 2011. Thompson Sampling for Dynamic Multi-Armed Bandits. In *Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops - Volume 01*. ICMLA '11 Washington, DC: IEEE Computer Society pp. 484–489. Accessed October 4, 2020, https://doi.org/10.1109/ICMLA.2011.144.

Hainmueller, Jens, and Dominik Hangartner. 2013. "Who Gets a Swiss Passport? A Natural Experiment in Immigrant

Discrimination." *American Political Science Review* 107(1): 159–87.

Khanna, Kabir, and Gaurav Sood. 2018. "Motivated Responding in Studies of Factual Learning." *Political Behavior* 40(1): 79–101.

Lei, Huitian, Ambuj Tewari, and Susan A. Murphy. 2017. "An Actor-Critic Contextual Bandit Algorithm for Personalized Mobile Health Interventions." arXiv e-prints arXiv:1706.09090v1 [stat.ML].

Li, Lihong, Wei Chu, John Langford, and Robert E. Schapire. 2010. "A Contextual-Bandit Approach to Personalized News Article Recommendation." arXiv e-prints arXiv:1003.0146v2 [cs.LG].

Locatelli, Andrea, Maurilio Gutzeit, and Alexandra Carpentier. 2016. "An Optimal Algorithm for the Thresholding Bandit Problem." arXiv e-prints arXiv:1605.08671v1 [stat.ML].

Lotze, Thomas, and Markus Loecher. 2014. *Bandit: Functions for Simple A/B Split Test and Multi-Armed Bandit Analysis.* R package version 0.5.0. https://CRAN.R-project.org/package=bandit

Ludwig, Jens, Jeffrey R. Kling, and Sendhil Mullainathan. 2011. "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives* 25(3): 17–38.

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. "How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It." *American Journal of Political Science* 62(3): 760–75.

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science*2:109–38.

Murphy, Susan A. 2003. "Optimal Dynamic Treatment Regimes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65(2): 331–55.

Nie, Xinkun, Xiaoying Tian, Jonathan Taylor, and James Zou. 2017. "Why Adaptively Collected Data Have Negative Bias and How to Correct for It." arXiv e-prints arXiv:1708.01977v2 [stat.ML].

Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115(2): 200–49.

Pallmann, Philip, Alun W. Bedding, Babak Choodari-Oskooei, Munyaradzi Dimairo, Laura Flight, Lisa V. Hampson, Jane Holmes, Adrian P. Mander, Matthew R. Sydes, Sofía S. Villar, et al. 2018. "Adaptive Designs in Clinical Trials: Why UseThem, and How to Run and Report Them." *BMC Medicine* 16(1): 29.

Prior, Markus, and Arthur Lupia. 2008. "Money, Time, and Political Knowledge: Distinguishing Quick Recall and Political Learning Skills." *American Journal of Political Science* 52(1): 169–83.

Prior, Markus, Gaurav Sood, Kabir Khanna, et al. 2015. "You Cannot Be Serious: The Impact of Accuracy Incentives on Partisan Bias in Reports of Economic Perceptions." *Quarterly Journal of Political Science* 10(4): 489–518.

Russo, Daniel, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. 2017. "A Tutorial on Thompson Sampling." arXiv e-prints arXiv:1707.02038v3 [cs.LG].

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press.

Sydes, Matthew R., Mahesh K. B. Parmar, Malcolm D. Mason, Noel W. Clarke, Claire Amos, John Anderson, Johann de Bono, David P. Dearnaley, John Dwyer, Charlene Green, Jordana Jovic, Alistair W. S. Ritchie, J. Martin Russell, Karen Sanders, George Thalmann, and Nicholas D. James. 2012. "Flexible Trial Design in Practice-Stopping Arms for Lack-of-Benefit and Adding Research Arms Mid-Trial in STAMPEDE: A Multi-Arm Multi-Stage Randomized Controlled Trial." *Trials* 13(1): 168.

Thompson, William R. 1933. "On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Aamples." *Biometrika* 25(3/4): 285–94.

Thompson, William R. 1935. "On the Theory of Apportionment." *American Journal of Mathematics* 57(2): 450–56.

Trippa, Lorenzo, Eudocia Q. Lee, Patrick Y. Wen, Tracy T. Batchelor, Timothy Cloughesy, Giovanni Parmigiani, and Brian M. Alexander. 2012. "Bayesian Adaptive Randomized Trial Design for Patients with Recurrent Glioblastoma." *Journal of Clinical Oncology* 30(26): 3258.

Urteaga, Iñigo, and Chris H. Wiggins. 2017. "Bayesian Bandits: Balancing the Exploration-Exploitation Tradeoff Via Double Sampling." arXiv e-prints arXiv:1709.03162v2 [stat.ML].

Villar, Sofía S., Jack Bowden, and James Wason. 2015. "Multi-Armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges." *Statistical Science* 30(2): 199.

Wason, James M.S., and Lorenzo Trippa. 2014. "A Comparison of Bayesian Adaptive Randomization and Multi-Stage Designs for Multi-Arm Clinical Trials." *Statistics in Medicine* 33(13): 2206–21.

Wason, James M.S., Peter Brocklehurst, and Christina Yap. 2019. "When to Keep It Simple—Adaptive Designs Are Not Always Useful." *BMC Medicine* 17(1): 1–7.

Wei, L.J.. 1988. "Exact Two-Sample Permutation Tests Based on the Randomized Play-the-Winner Rule." *Biometrika* 75(3): 603–6.

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *The Quarterly Journal of Economics* 134(2): 557–98.

Zhang, Kelly W., Lucas Janson, and Susan A. Murphy. 2020. "Inference for Batched Bandits." arXiv e-prints arXiv:2002.03217v2 [cs.LG].

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A: Algorithms
Appendix B: Estimation and theory
Appendix C: Worked Examples
Appendix D: Additional Simulations
Appendix E: Additional Information, Study One
Appendix F: Additional Analyses, Study Two
Appendix G: Study Three: An Adaptive Conjoint Trial