Adaptive Experimental Design: Prospects and Applications in Political Science

Molly Offer-Westort[*], Alexander Coppock[†], and Donald P. Green[‡]

August 24, 2018
Prepared for presentation at the Annual Meeting of the American Political Science Association, Boston, August 30th - September 2, 2018.

Abstract

Experimental researchers in political science frequently face the following inference problem: Which of several treatment arms produces the greatest return (where returns may be expressed in terms of campaign donations, new supporters of a political cause, adherents to a policy, etc.)? Multi-arm trials are typically conducted using a static design in which fixed proportions of the subject pool are allocated to each arm. However, a growing statistical literature suggests that adaptive experimental designs may be far more efficient in finding the most effective treatment arm. An important class of adaptive designs uses probability matching strategies to dynamically allocate subjects to treatment arms. We review the underlying assumptions of the multi-arm bandit framework and suggest that it has many potential applications in political science. We discuss the design and analysis of original experiments using this approach and compare their efficiency to a more traditional static design.

[*]Molly Offer-Westort is a Ph.D. Candidate in the Department of Political Science, Yale University
[†]Alexander Coppock is Assistant Professor of Political Science, Yale University.
[‡]Donald P. Green is the Burgess Professor of Political Science, Columbia University

Experimentation in the social sciences often boils down to a search for the intervention that maximizes some desired outcome. What pricing strategy maximizes demand for vaccines in low-income countries? (Cook et al. 2009) Which of the many ways of monitoring corruption among public officials minimizes the amount 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 et al. 2011).

Experiments that assess the relative effectiveness of competing interventions, be they policies or messages, confront a fundamental problem: the list of interventions under consideration is so long that it is prohibitively costly and time-consuming to test more than a small subset. Furthermore, even if monetary costs were no object, there remains an ethical concern that 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.

Adaptive trails (Chin 2016; Chow and Chang 2008) represent a design-based attempt to increase the speed and efficiency with which multi-arm trials discern the best-performing intervention or interventions. In contrast to conventional static designs, which allocate a fixed proportion of subjects to each arm throughout the trial, adaptive trials adjust the allocation as the trial unfolds, investing an ever-larger share of the subject pool in more promising treatment arms.[1] Adaptive trails are most likely to pick the true winner when the best arm is substantially better than its competitors; in such cases the design often declares a winner with much greater confidence than would have been the case under a static design.

That said, adaptive designs are no panacea. In situations where several treatment arms are equally effective (or nearly so), adaptive algorithms may equivocate, allocating more subjects to better performing arms whose initial success was due only to sampling fluctuation. In the end, there is no guarantee that the researcher will discover the truly best intervention and no guarantee that the adaptive design will have allocated more subjects to the truly best option than the static design. 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.

[1] Adaptive trials encompass a broad class of designs that potentially evolve based on interim results. Here we focus exclusively on the allocation of subjects, but other adaptive design adjustments include changing the treatments or halting the trial entirely. See Chin (2016).

The aim of this paper is to introduce political scientists to adaptive designs, highlighting the conditions under which they do or do not outperform conventional static designs. We begin by introducing the basic features of one commonly used adaptive design, known in the literature as Thompson Sampling, or more generally randomized probability matching (Thompson 1933, 1935). The features of this design are illustrated through simulation of multi-arm trials, some of which are more favorable to adaptive designs, while others expose their limitations. Next, we turn to empirical applications involving the wording of ballot measures. We first present a pair of multi-arm trials. The first assembles actual ballot measures proposing changes in the minimum wage; we conduct an adaptive trial to discern which wording maximizes voter support. The second experiment uses the same design to maximize voter support for right to work proposals. These two examples illustrate the conditions under which adaptive designs work well. In the case of right to work proposals, there is a clear winner that adaptive design identifies quickly and with a high degree of statistical precision. Results are more ambiguous for minimum wage proposals, where several proposals seem equally promising. We conclude by discussing another empirical example, this one conducted in collaboration with an interest group seeking to isolate the most popular way of wording a ballot measure, where the set of feasible alternatives includes thousands of possibilities. We discuss strategies for conducting adaptive trials when the number of potential interventions is large relative to the number of subjects.

Adaptive Trials

In substantive domains from advertising to biomedical research, adaptive trials are used to speed the search for the best-performing intervention. One of the central ideas in adaptive trial design is Thompson sampling, a heuristic approach to the task of selecting from among multiple arms with the objective of maximizing reward. Thompson sampling randomly allocates subjects to treatment arms according to their probability of being the "best," i.e., returning highest reward. We will focus primarily on the Thompson samplings algorithm here, although there are many alternatives, such as the Greedy algorithm, which selects the arm with the highest empirical mean, and the upper confidence bound (UCB) algorithm, which selects the arm with the highest upper bound on the confidence interval around its sample mean. Allocation rules vary across performance metrics, such as statistical power, type-I error rates, and bias, and the appropriate rule will depend on the time-horizon (Villar et al. 2015).

The typical tradeoff addressed in such settings is between exploration, testing arms to gather information about them, and exploitation, selecting the most promising arm(s). Experimenters would like to gain more information about the probability of success of each arm so that they can be confident in selecting the best arm or arms. However, over-exploring could mean wasting draws on under-performing arms. On the other hand, too quickly exploiting arms that perform relatively well, without having fully explored the available options, could mean ignoring potentially superior arms. Thompson sampling provides an intuitive manner of trading off exploration and exploitation: when the researcher does not have much information about which arm is the best, the algorithm will facilitate exploration; as more information is gained, the best-performing arms are increasingly exploited. Which arm is best is calculated as a function of the payoff, or reward the experimenters get as a consequence of selecting the arm.

While this approach is generalizable to continuous rewards, we consider the application to binary rewards, where each observation is either a success or a failure. Here k arms have an unknown probability of success $(p_1, \ldots, p_k)$, following the respective Binomial distributions. An experimenter assigns some prior to the probability of success of each arm; in practice this is generally taken to be $U(0, 1)$. The posterior then follows a Beta distribution with α equal to 1 plus the total number of successes observed for that arm, and β equal to 1 plus the total number of failures observed from that arm. In each period, arms are randomly selected according to their probability of being the optimal arm,[2] and rewards are observed; in the first period, all arms have equal probability of being the best, and so all arms are sampled with equal probability. At the end of the period, the posterior is updated according to the successes and failures in that period, the probability that each arm is the best is re-calculated, and sampling continues in the next period.

Thompson sampling can be adapted to account for settings in which there is drift in parameter values over time (Gupta et al. 2011), or can be contextualized based on covariate values (Agrawal and Goyal 2013). In some applications, such as the ones considered here, adaptive trials end after a fixed period or when a pre-determined number of subjects have participated in the trial. In other applications, the trial stops when the best-performing arm achieves a specified posterior probability; when used to establish statistical significance of effects, such stopping rules can exacerbate false discovery, as the trial may stop if the best-performing arm surpasses the target due to chance (Berman et al. 2018).

[2] We estimate this value through simulation, taking a series of random draws from the posterior probability distributions of each arm, and calculating the share of the series in which each arm had the highest draw, as implemented in the `bandit` package.

Here, the objective is to maximize reward, not, as is common in the social sciences, to estimate the difference between treatment and control. Indeed, if the "best" treatment arm has a much higher probability of success than the control arm, the control arm will be assigned a relatively small sample, in which case the difference in means will generally have a larger standard error than under balanced assignment. To improve power in such settings, Villar et al. (2015) propose a composite design, where treatments are allocated adaptively, but in which a set portion of patients are allocated to the control group. Additionally, Nie et al. (2017) demonstrate that under certain common conditions, sample means from adaptive experiments are systematically negatively biased. Ex-post approaches to estimation to reduce bias in such settings, such as inverse propensity score methods, can exhibit large variance (Dimmery 2018; Nie et al. 2017). Consequently, if an unbiased estimate of treatment effects is the primary objective of a study, standard static designs may be preferable. Alternatively, procedures to reduce bias may be integrated into the design stage of adaptive experiments. A static experiment may be conducted following, or alongside, an adaptive experiment to produce unbiased treatment effect estimates; or Nie et al. (2017) propose a novel randomization algorithm founded on selective inference methods, which may have lower RMSE compared to data splitting designs with comparable sample sizes.

We consider an additional setting where arms are structured, composed of factorial components, with each component contributing to the success rate of the arm. Combinations of factors can quickly result in a large number of arms that may be unwieldy for exploration. In the case of binary rewards, the reward distribution can be modeled by a probit regression on the respective factorial components (see, e.g., Scott 2010; for more general cases, see Filippi et al. 2010). Modeling assumptions allow us to pool information across arms sharing common components and to estimate success probabilities for arms which we have not observed. Such model-based approaches can be used to select from among a large number of candidate profiles. We conducted an additional confirmatory stage to serve as a run-off for the arms with the highest predicted probability of success.

Simulations Illustrating How Adaptive Designs Work

We illustrate the method with simulations in several settings. In each case, there are nine arms, each with some true probability of success set according to $(p_1, \ldots, p_9)$. In the first period, we assume a uniform distribution over the probability of success for each of the arms. As the probability of being the best is equal for all of the arms prior to the start of

the experiment, each arm is sampled with equal probability in the first period. We sample 100 observations for each of 14 periods, updating each arms' estimated probability of success and probability of being the best after each period, and sampling in the subsequent period accordingly. The choice of 14 periods is arbitrary but anticipates an empirical example presented below, which runs for 14 days.

In the first case, one arm has a true 0.2 probability of success, and the remaining eight arms have a 0.1 probability of success. Within two periods, the true best arm (presented as a solid orange line) takes a clear lead in probability of being the best, shown in the left facet of Figure 1. By the end of the 14-period experiment, the true best arm is assigned nearly 90% probability of being the best. In the right facet we plot potential value remaining, or per-play regret, a measure of how much success rates might be improved by switching to another arm (Scott 2015). By the end of the experiment, we predict that we could improve success rate over the arm ranked as being the best by as much as 9%. Since the amount of improvement that could be expected to be achieved with additional experimentation is relatively small, the experimenter may satisfice with this outcome.

In the second case, the best arm has only a 0.12 true probability of success, compared to a 0.1 probability of success for the remaining 8 arms. The 14 period experiment does not allow enough time to come to a clear conclusion about which arm is the best. Indeed, in our illustrative example we assign the best arm (again in solid orange) only 9.3% probability of being the best, whereas we assign an inferior arm 30% probability of being the best, shown in the left facet of Figure 2. Value remaining is also relatively high; by the end of the experiment, we predict that we could improve success rate over the arm ranked as being the best by as much as 82%.

Finally, we consider a case where the best arm has a 0.2 true probability of success, a second-best arm has a 0.18 probability of success, and remaining arms have 0.1 probability of success. The second best arm (in dotted green) quickly takes a lead, but by the end of the experiment we assign the true best arm a 50% probability of being the best, shown in the left facet of Figure 3. In such cases, the true best arm may end up assigned the highest probability of being the best, but when an arm underperforms by chance, the consequent low sampling probability means that our estimates of the arm's success might not recover. While we pick the correct arm in this case, the moderate potential value remaining at the end of the experiment indicates that an alternative arm may could surpass the success rate of the winning arm by as much as 29%.

To compare the performance of adaptive trials in the three scenarios more rigorously, we

Figure 1: Case 1, clear winner

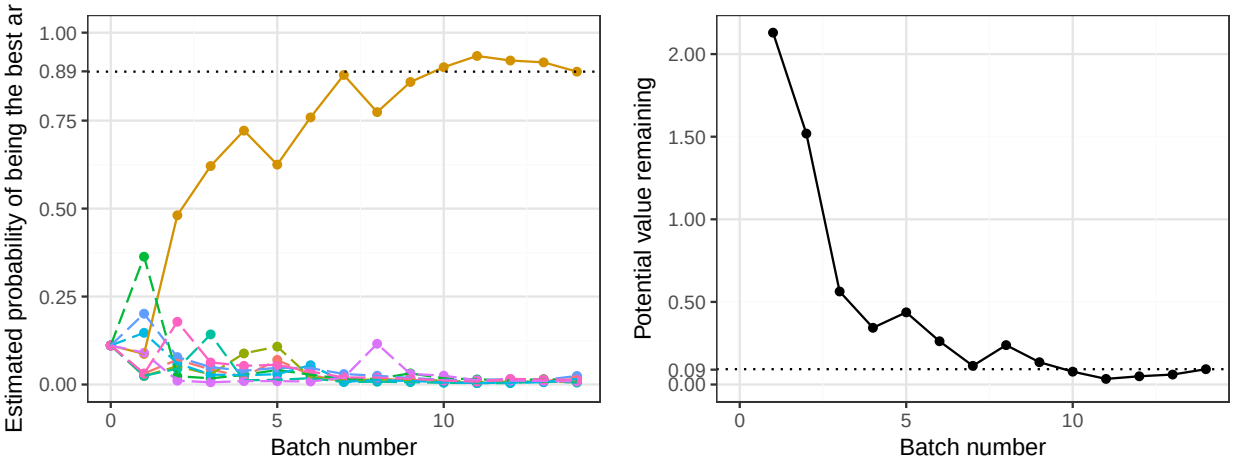


Figure 2: Case 2, no clear winner

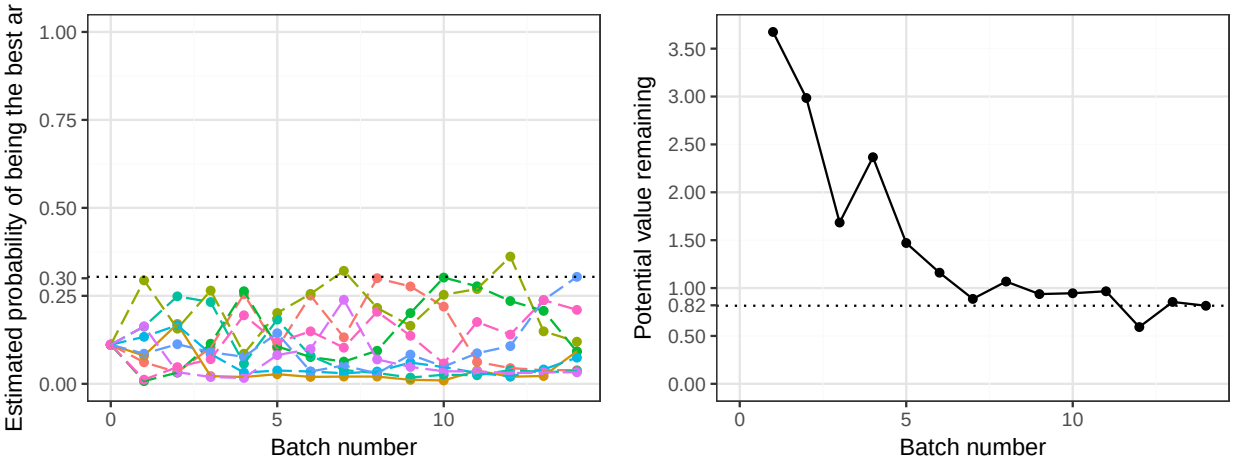
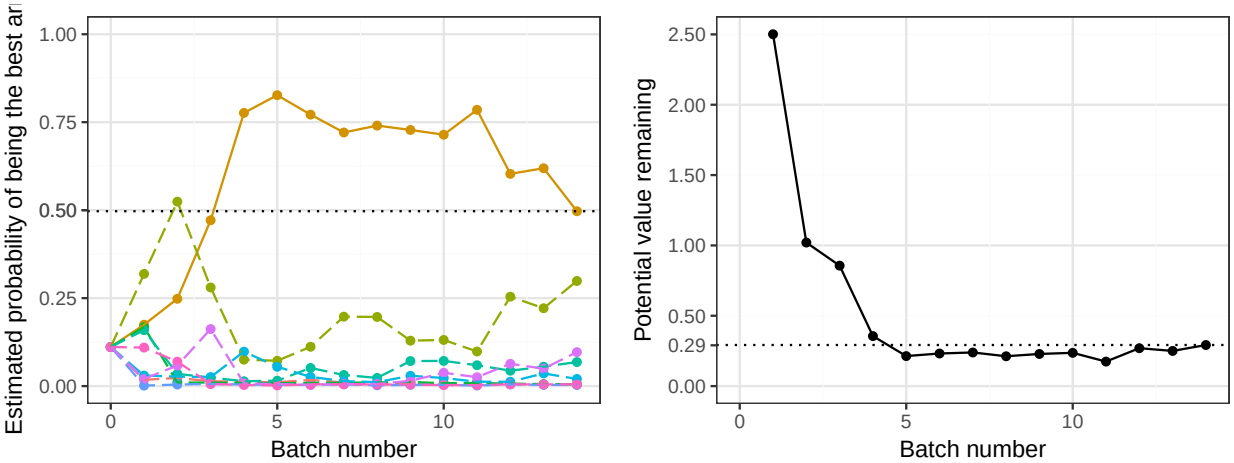
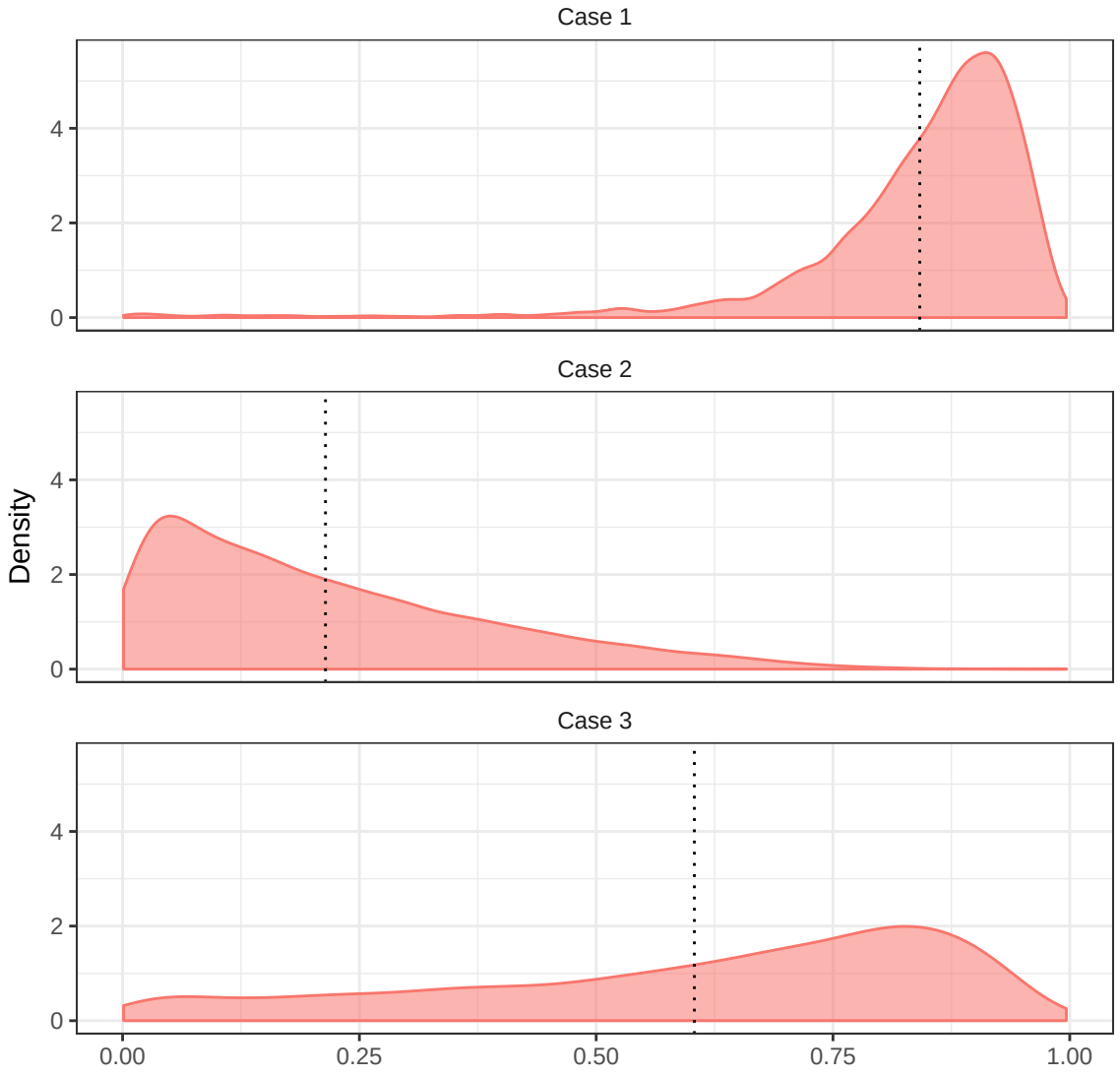


Figure 3: Case 3, competing winner



plot the distributions of predicted probability of being the best for the true best arm across 5,000 simulations in Figure 4. In the first scenario, by the end of the experiment we assign 84.2% probability to the true best arm on average, and pick the true best arm in 98.9% of trials. In the second case, we only assign 21.4% probability to the true best arm, and select it in only 34.5% of trials. In the third case, we assign the true best arm 60.4% probability of being the best, and select it in 75.9% of trials. The bottom line seems to be that adaptive trials work well when there is a clear winner to be found or when a small subset of treatment arms stand above the rest.

Figure 4: Distribution of probabilities of being the best for true best arm



Study 1: Two Multi-arm Adaptive Trials

We recruited 1100 subjects from Amazon's Mechanical Turk (MTurk) to participate in Study 1. Convenience samples obtained on Mechanical Turk are far from representative of the national population, but do provide a fertile testing ground for experimental studies. Recent studies have revealed a close correspondence of experimental estimates obtained on MTurk and probability samples (Mullinix et al. 2015; Coppock 2017; Coppock et al. 2018). Our study ran from June 21st, 2018 to June 30th, 2018. We paid respondents $1.00 each for their participation.

Design

After answering a series of demographic questions, all subjects rated two ballot measures, one on the minimum wage and the second on right to work. We adapted the wording of these measures from real proposals, making only small wording changes to facilitate consistency of measurement across arms. We implemented a composite design parallel to the controlled Gittins approach recommended by Villar et al. (2015); for each type of ballot measure 90% of treatments were assigned according to Thompson Sampling, and 10% were assigned under simple random assignment, with equal probability for each treatment.

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.[3] The 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." For both rating tasks, the outcome question was asked, "If this measure were on the ballot in your state, would you vote in favor or against?" The full text of all treatments is presented in Table 1.

Results

We present two sets of results. The first, presented in Figure 5, is the over time development of the posterior probability that each arm is the best. The second, presented in Figure 6, is a straightforward comparison of the average approval of each proposal. A key feature of

[3]We obtained the current minimum wage rate from `https://en.wikipedia.org/wiki/Minimum_wage_in_the_United_States`. These rates are included in the online appendix. For states that do not have a minimum wage, we imputed the federal minimum wage value.

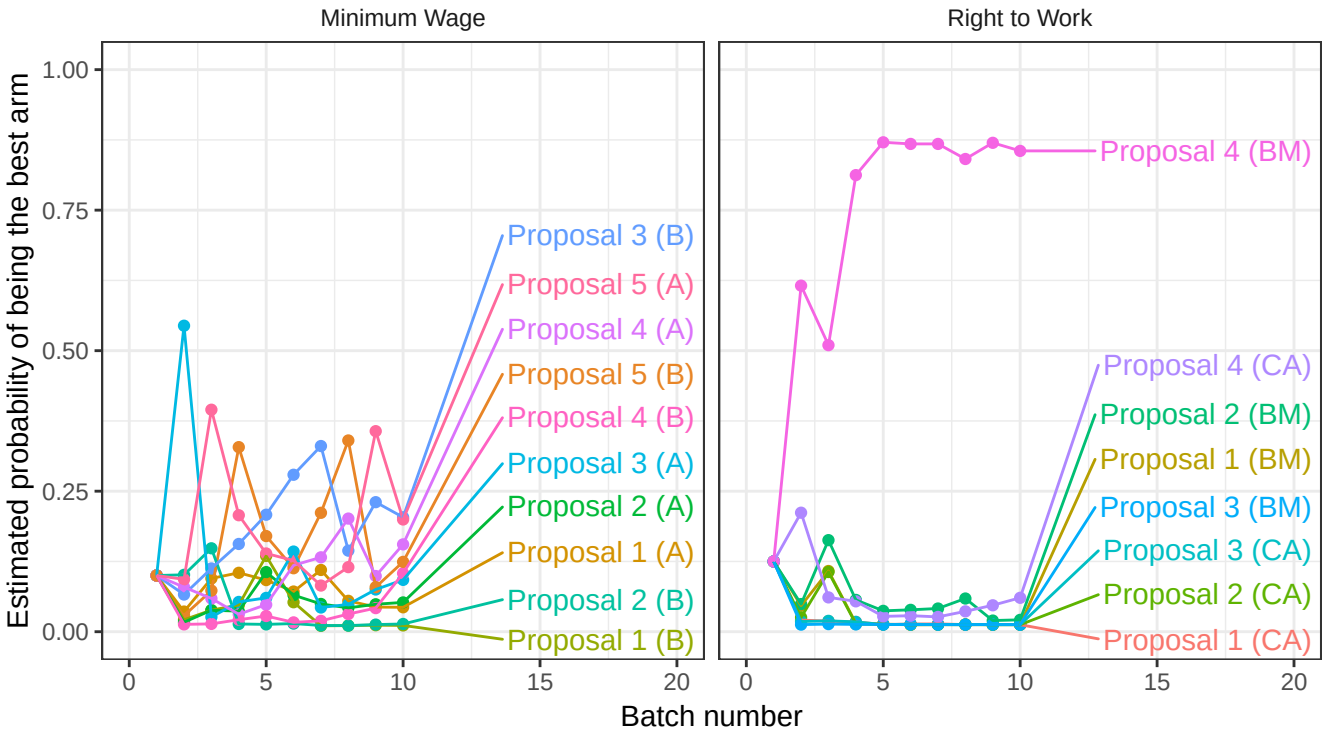
Table 1: Treatments and Outcome Measures

	Minimum Wage	Right to Work
Question Text	Imagine that the following ballot measure were up for a vote in your state. The measure would: [**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. The measure would [**amend the State Constitution to**]: [**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	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.	prohibit, as a condition of employment, forced membership in a labor organization (union) or forced payments of dues or fees, in full or pro-rata ("fair-share"), to a union. The measure will also make any activity which 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	raise the minimum wage [**from {current}**] to {current + 1} per hour effective September 30th, 2021. Each September 30th thereafter, minimum wage shall increase by $1.00 per hour until the minimum wage reaches {current + 5} per hour on September 30th, 2026. From that point forward, future minimum wage increases shall revert to being adjusted annually for inflation starting September 30th, 2027.	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	Shall the minimum wage for adults over the age of 18 be raised [**from {current}**] to {current + 1} per hour by January 1, 2019?	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	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.	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	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.	

Boldface text indicates randomly varied elements.

adaptive designs is that the probability of assignment to each condition varies over time. A default analytic approach when the probabilities of assignment are different for different units is to weight each observation by the inverse of the probability of assignment to the condition that it is in (see Gerber and Green (2012, chp. 4) for a textbook introduction to inverse probability weighting, or IPW). In this case, because we do not incorporate the changing probabilities of assignment into the procedure for estimating the probability that each arm is the best, we also present unweighted group averages. The resulting group averages are unbiased under the assumption that the potential outcomes of subjects are equivalent (in expectation) for each day of the experiment. For completeness, we also present group means that use IPW in the appendix. Because of the extreme weights, the ranking of the treatments is mildly different when we use IPW.

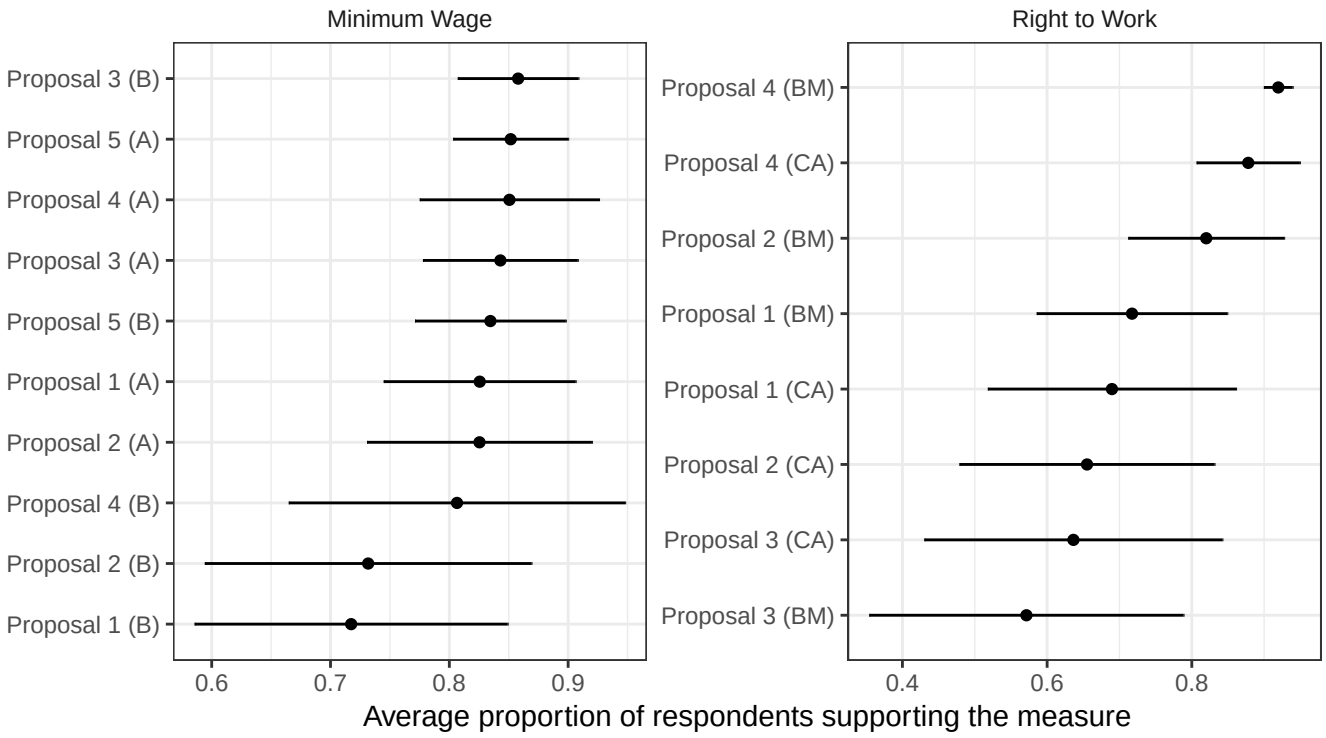
Figure 5: Study 1: Overtime Posterior Probabilities



Posterior probabilities updated after each day's data collection according to the algorithm described in footnote 2. "A" versions of the minimum wage proposals include the current minimum wage and "B" versions do not. "CA" versions of the right to work proposals are describes as "constitutional amendments" and "BM" versions are not.

The minimum wage study yielded no clear winner. The winning arm, by a hair, was

Figure 6: Study 1: Group Means



Group means are unweighted. "A" versions of the minimum wage proposals include the current minimum wage and "B" versions do not. "CA" versions of the right to work proposals are describes as "constitutional amendments" and "BM" versions are not.

Proposal 3 (B, without current minimum wage), with a posterior probability of being the best of 20.8% and a raw success rate of 85.8% over 183 trials. This arm was closely followed by Proposal 5 (A, with current minimum wage), Proposal 4 (A, with current minimum wage), and Proposal 3 (A, with current minimum wage). Out of 10 arms, only two had raw success rates under 80%; with similarly high probabilities of success across several arms, the best arm was not easily distinguishable.

By contrast, the right to work experiment immediately produced a standout arm that proved to be very successful throughout the study. Proposal 4 (framed as a ballot measure) ended with an 82.7% posterior probability of being the best arm, with a raw success rate of 92.0% over 183 trials. The second-best arm was also Proposal 4 (framed as a constitutional amendment) with a posterior probability of being the best of 9.2% and a raw success rate of 87.8% over 82 trials. This example underlines that even a small (4.2 percentage point) difference in success rates can translate into a very large (82.8 percentage point) difference in the probability of being the best 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 100 times each, and each of the 8 arms in the right to work experiment 125 times each. We conduct simulations in which we use our estimated success rates as the truth. In simulations of the minimum wage experiment, we picked the winning Proposal 3 (B) 25% of the time, and, on average, assigned this arm a posterior probability of being the best of 21%. Conversely, for the right to work experiment, we picked the winning Proposal 4 (BM) 83% of the time, and, on average, assigned this arm a posterior probability of being the best of 74%.

Considering Figure 6, we note that a feature of the adaptive design is that the proposals with the highest raw success rates also have the tightest standard errors, as these arms tend to receive more samples than arms with lower success rates. This is appropriate when the performance of the best arm is considered a priority by the researcher and success rates of poorly-performing arms is not of particular interest. Under a static design, our certainty regarding the best arm would also have been much less exact. For the minimum wage experiment, our standard errors around our estimate of probability of success for the best arm in a static design would have been 135% as large as they were under our realization of the adaptive design, assuming the realized success rate is truth. For the right to work experiment, standard errors on the probability of success of the best arm would have been 240% as large as our realization. The adaptive design appears to offer clear advantages in

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

Study 2: An Adaptive Conjoint Trial

Study 2 extends the adaptive design beyond the multi-arm design (as in Study 1) to factorial designs, often referred to a conjoint experiments.[4] Like any factorial design, conjoint designs enable the study of multiple dimensions of preference within the same experiment. However, even a moderately complex conjoint experiment with, for example, five attributes with four levels each would produce $4^5 = 1024$ possible treatment configurations. Even with an adaptive design, discovering the best arm among such a vast number of possibilities by brute force would be prohibitively expensive.

Our solution is to take advantage of the factorial design to put structure on the problem. As a first approximation, we generate an *additive* model of being the best arm. The model is additive in the sense our prediction of the probability of each arm being the best is the *sum* of the Average Marginal Component Effects (AMCEs) of each component in the profile. We used these probabilities in order to perform our adaptations. We modeled probability of success as a probit regression on the components, using the `R` package `MCMCpack`, assigning an improper uniform prior to the coefficients.

Design

For study 2, we recruited approximately 4,000 subjects over 14 days on Lucid, a marketplace for online survey respondents. Like Mechanical Turk, Lucid provides respondents who are not representative of the national population, although demographic diversity is greater on Lucid than on MTurk (Coppock and McClellan 2018).

Subjects rated two ballot proposals each. Unlike Study 1, these ballot proposals were created in collaboration with [redacted], an organization promoting a ballot measure on the subject of [redacted], in the state of [redacted]. Each proposal consisted of a title, and four bullet points. Proposals were assigned four out of eight potential content factors, with each bullet point consisting of one level of an assigned factor. Table 2 lists number of levels in each factor. The total number of possible ballot measures is 3,388. Because the results of

[4]Conjoint experiments were introduced to political science by Hainmueller et al. (2014). For an application of the conjoint design to the study of candidate partisanship on preferences over mayoral candidate attributes, see Kirkland and Coppock (2018).

this study may inform choices by our partner organization in advance of the 2018 elections, we omit the content of this study for the present time.

Table 2: Factor composition

Factor	Arms
Title	2 levels
Factor 1	2 levels
Factor 2	2 levels
Factor 3	3 levels
Factor 4	1 level
Factor 5	2 levels
Factor 6	9 levels
Factor 7	1 level
Factor 8	1 level

Results

We present three sets of results.

First, we trace the development of the 50 arms that had the highest posterior probability of being the best after the 11^{th} day, by which point 2,961 subjects had rated 5,922 profiles. Figure 7 shows progression of predicted probability of being the best over these 11 days. Posterior probabilities evolved considerably from the start of the trial, when every arm had the same 0.03% probability of begin best. Nevertheless, no arm exceeded a 5% chance of being the best arm after day 11.

Second, we analyze our experiment as in a traditional conjoint. Figure 8 shows the estimated AMCEs, calculated via OLS with standard errors clustered by subject. [5] Because we cannot share the details of what each factor refers to, Figure 8 serves only to demonstrate the feasibility of analyzing data from an adaptive conjoint in the standard way.

Finally, we validated our model by running a static "bake-off" trial among eight arms that were predicted to have a high probability of success. We conducted this eight-arm trial among 971 subjects also drawn from Lucid, who rated 1,942 profiles over the remaining days

[5] In the first round of the experiment, treatment was allocated across all arms under uniform random assignment. Due to the large number of possible factor combinations and chance assignment, the single level in factor 8 was collinear with other combinations of factors, and so effects were not estimated for this factor. Subsequently, profiles with factor 8 were assigned with very low probability, and so we are not able to estimate effects for this factor. Issues of potential collinearity could be addressed with a fractional factorial design or otherwise enforced balance across factors in randomization in the first period. We present effect estimates here for levels excluding factor 8.

Figure 7: Top 50 finishing arms in adaptive conjoint, over time

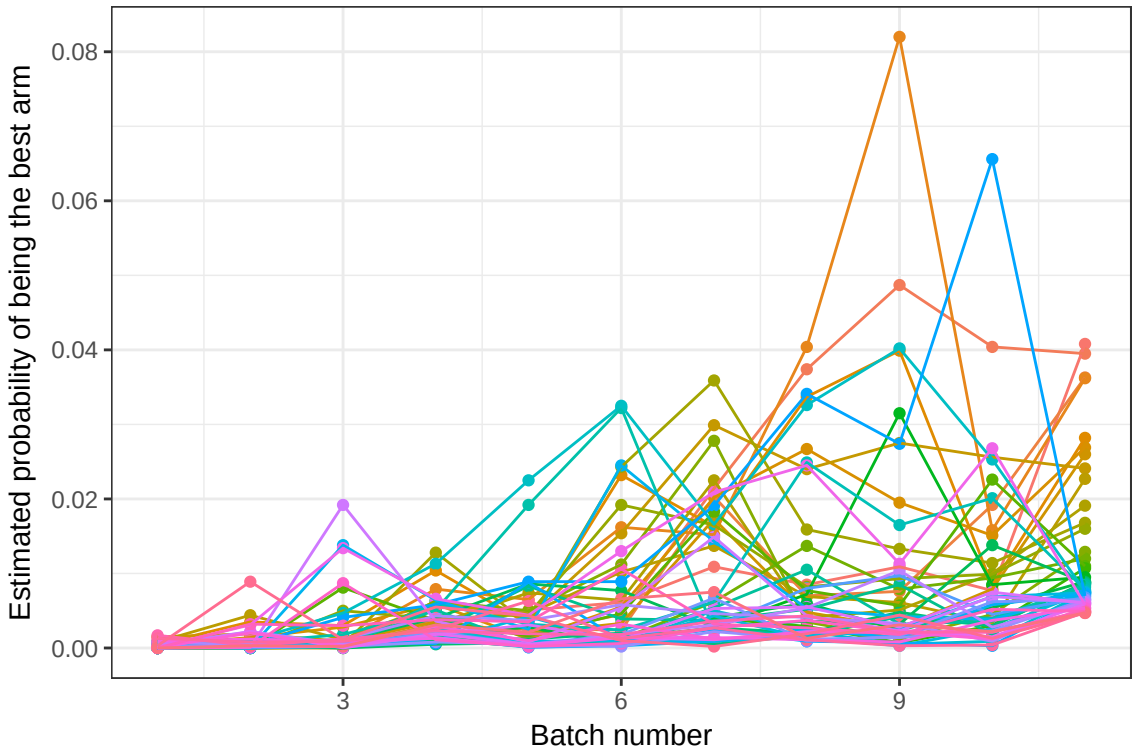
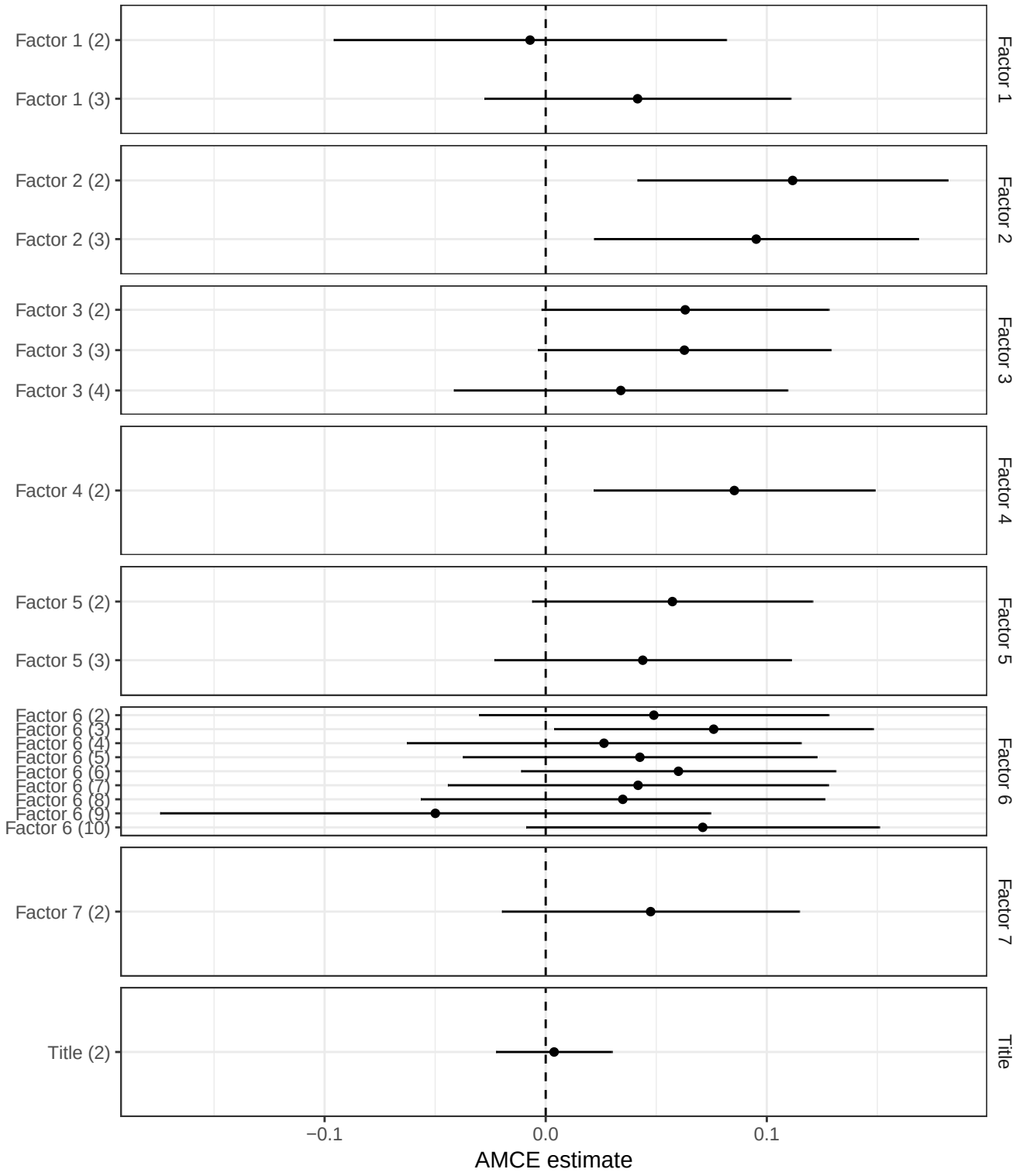


Figure 8: Average Marginal Component Effects



in the experiment. The eight arms were selected by picking the top two arms according to four separate analytic strategies:

1. A probit regression of the outcome on the levels of each factor with no interactions (a "main effects" model).

2. A probit regression of the outcome on each factor plus all two-way interactions of the factors.

3. An elastic net model with all factors and their two-way interactions[6]

4. A Bayesian Additive Regression Trees (BART) model (Chipman et al. 2010; Hill 2011; Green and Kern 2012) that flexibly accommodates interactions of any order among all factors.

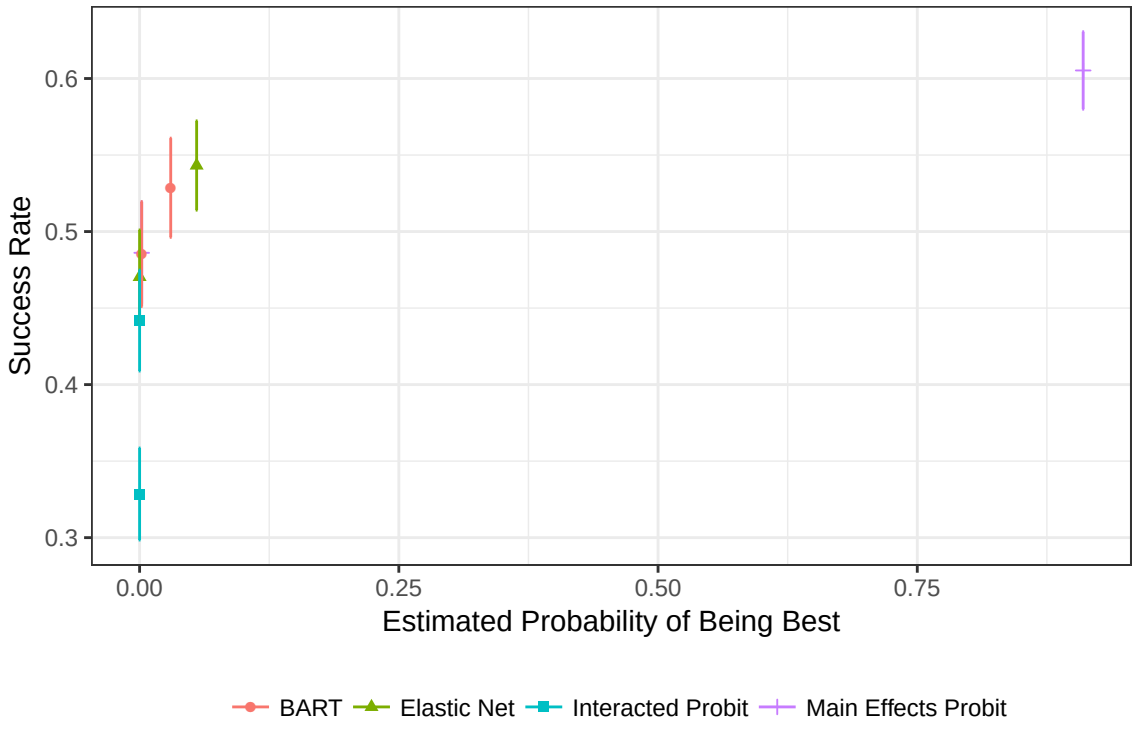
Figure 9 shows the results of this additional study. Perhaps surprisingly, the least complex model (main effects probit) was by far the most effective in choosing the winner of the bake-off. Interacted probit performed the worst, with the two models that allow for some regularization (elastic nets and BART) performing somewhere in the middle. Substantively, this pattern of results implies that when evaluating this ballot measure, respondents considered each of the elements in isolation and their overall evaluation was approximately the sum of their element-by-element evaluations.

Discussion

The growth and development of experimentation in the social sciences has led to increasing sophistication in the design of multi-arm 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 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

[6]We implement k-fold cross-validation under the `R glmnet` package, with $\alpha = 0.5$, using λ that gives minimum cross-validated error.

Figure 9: 8-arm bake-off



to work ballot measure was estimated with a standard error that was 42% as large as would have been the case under a static design.

The adaptive allocation of subjects, however, is of little 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 and reallocates the sample to obtain a somewhat more precise assessment of the winning arm's payoff. In this application, an adaptive design still outperformed the expected outcome from a static design in terms of the precision with which the winning arm's payoff was estimated, but the gains were far less dramatic than the right to work application. Researchers considering the use of adaptive sample allocation should therefore reflect on their prior beliefs about the effectiveness of the treatment arms they plan to investigate. The more variable the effects, the more valuable an adaptive design is likely to be.

This point also holds for studies in which the aim is to compare treatment arms to an untreated control group. Consider a hybrid design in which a static allocation is made to an untreated control group throughout the trial, but sample is allocated adaptively to the treatment arms. The adaptive component of the design aims to locate the best performing treatment arm, but the static 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.

An important research frontier is the efficient allocation of sample in the context of factorial designs. To our knowledge, ours is the first paper to consider adaptive design in the context of conjoint experiments in the social sciences, where the research aim is to find the combination of traits with the highest payoff. 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 to reduce the set of promising treatment combinations. In our application, we addressed this challenge by way of a two-part tournament: an initial phase in which adaptive allocation sought to identify effective treatments and a bake-off phase in which the two finalists proposed by each of four statistical models competed in a further adaptive trial. This procedure is by no means the only way to conduct a tournament of this kind, but it remains noteworthy

that in this application a nominee proposed by a relatively simple additive model easily outdistanced its competitors. Further empirical tests are needed in order to assess whether additive models perform well in other substantive domains and whether the performance of more sophisticated models could be improved by restructuring the initial search phase to more reliably explore interactions among factors.

References

Agrawal, Shipra and Navin Goyal. 2013. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*. pp. 127–135.

Berman, Ron, Leonid Pekelis, Aisling Scott and Christophe Van den Bulte. 2018. "p-Hacking and False Discovery in A/B Testing.".

Chin, Richard. 2016. *Adaptive and flexible clinical trials*. CRC Press.

Chipman, Hugh A., Edward I. George and Robert E. McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 4(1):266–298.

Chow, Shein-Chung and Mark Chang. 2008. "Adaptive design methods in clinical trials–a review." *Orphanet journal of rare diseases* 3(1):11.
URL: *http://doi.org/10.1186/1750-1172-3-11*

Cook, Joseph, Marc Jeuland, Brian Maskery, Donald Lauria, Dipika Sur, John Clemens and Dale Whittington. 2009. "Using private demand studies to calculate socially optimal vaccine subsidies in developing countries." *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management* 28(1):6–28.

Coppock, Alexander. 2017. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* . Forthcoming.

Coppock, Alexander and Oliver A. McClellan. 2018. "Validating the Demographic, Political, Psychological, and Experimental Results Obtained from a New Source of Online Survey Respondents." *Unpublished manuscript* .

Coppock, Alexander, Thomas J. Leeper and Kevin J. Mullinix. 2018. "The Generalizability of Heterogeneous Treatment Effect Estimates Across Samples." Unpublished manuscript.

Dimmery, Drew. 2018. "Adaptive Experimental Design for Social Science." Manuscript in preparation.

Filippi, Sarah, Olivier Cappe, Aurélien Garivier and Csaba Szepesvári. 2010. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*. pp. 586–594.

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

Green, Donald P. and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3):491–511.

Gupta, Neha, Ole-Christoffer Granmo and Ashok Agrawala. 2011. Thompson sampling for dynamic multi-armed bandits. In *2011 10th International Conference on Machine Learning and Applications Workshops*. IEEE pp. 484–489.

Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22(1):1–30.
URL: *http://pan.oxfordjournals.org/cgi/doi/10.1093/pan/mpt024*

Hainmueller, Jens and Dominik Hangartner. 2013. "Who gets a Swiss passport? A natural experiment in immigrant discrimination." *American political science review* 107(1):159–187.

Hill, Jennifer L. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1):217–240.

Kirkland, Patricia A. and Alexander Coppock. 2018. "Candidate Choice Without Party Labels:." *Political Behavior* 40(3):571–591.

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

Martin, Andrew, Kevin Quinn and Jong Hee Park. 2011. "MCMCpack: Markov Chain Monte Carlo in R." *Journal of Statistical Software, Articles* 42(9):1–21.
URL: *https://www.jstatsoft.org/v042/i09*

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2:109–138.

Nie, Xinkun, Xiaoying Tian, Jonathan Taylor and James Zou. 2017. "Why adaptively collected data have negative bias and how to correct for it." *arXiv preprint arXiv:1708.01977* .

Olken, Benjamin A. 2007. "Monitoring corruption: evidence from a field experiment in Indonesia." *Journal of political Economy* 115(2):200–249.

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

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

Simon, Noah, Jerome Friedman, Trevor Hastie and Rob Tibshirani. 2011. "Regularization paths for Cox's proportional hazards model via coordinate descent." *Journal of statistical software* 39(5):1.

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

Thompson, William R. 1935. "On the theory of apportionment." *American Journal of Mathematics* 57(2):450–456.

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: a review journal of the Institute of Mathematical Statistics* 30(2):199.

A Minimum Wage Rates

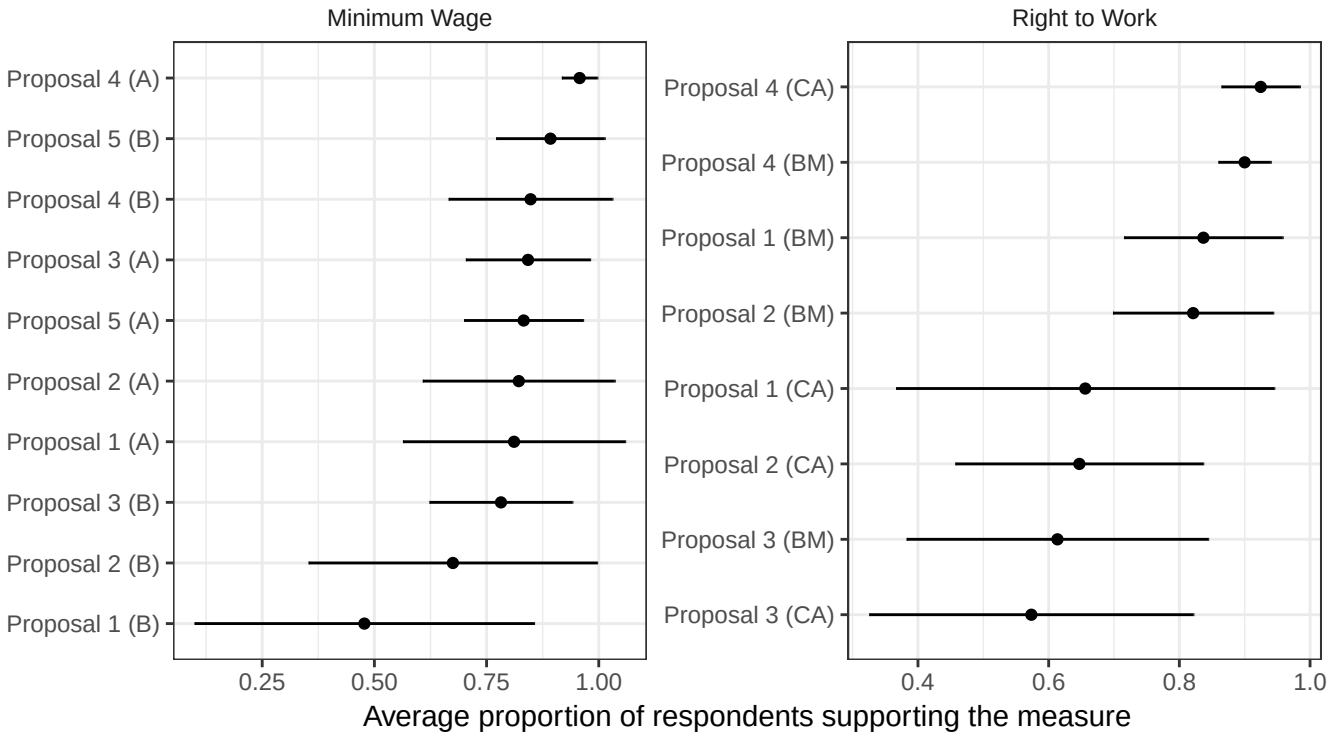
Table A.3: Minimum Wage rates as of June, 2018

State	Minimum Wage
Alabama	$7.25
Alaska	$9.84
Arizona	$10.50
Arkansas	$8.50
California	$11.00
Colorado	$10.20
Connecticut	$10.10
Delaware	$8.25
Florida	$8.25
Georgia	$7.25
Hawaii	$10.10
Idaho	$7.25
Illinois	$8.25
Indiana	$7.25
Iowa	$7.25
Kansas	$7.25
Kentucky	$7.25
Louisiana	$7.25
Maine	$10.00
Maryland	$9.25
Massachusetts	$11.00
Michigan	$9.25
Minnesota	$9.65
Mississippi	$7.25
Missouri	$7.85
Montana	$8.30
Nebraska	$9.00
Nevada	$8.25
New Hampshire	$7.25
New Jersey	$8.60
New Mexico	$7.50
New York	$10.40
North Carolina	$7.25
North Dakota	$7.25
Ohio	$8.30
Oklahoma	$7.25
Oregon	$10.25
Pennsylvania	$7.25
Rhode Island	$10.10
South Carolina	$7.25
South Dakota	$8.85
Tennessee	$7.25
Texas	$7.25
Utah	$7.25
Vermont	$10.50
Virginia	$7.25
Washington	$11.50
West Virginia	$8.75
Wisconsin	$7.25
Wyoming	$7.25

Source: `https://en.wikipedia.org/wiki/Minimum_wage_in_the_United_States`

B Additional Analyses

Figure B.10: Study 1: Group Means (IPW)



Group means are weighted by inverse probability weights. "A" versions of the minimum wage proposals include the current minimum wage and "B" versions do not. "CA" versions of the right to work proposals are describes as "constitutional amendments" and "BM" versions are not.

Table B.4: Adaptive conjoint MCMC probit model

	Factors
Title 0	−0.684
	(0.296)
Title 1	0.009
	(0.033)
Factor1.Level1	−0.021
	(0.114)
Factor1.Level2	0.105
	(0.088)
Factor2.Level1	0.281
	(0.088)
Factor2.Level2	0.240
	(0.092)
Factor3.Level1	0.160
	(0.084)
Factor3.Level2	0.160
	(0.085)
Factor3.Level3	0.085
	(0.095)
Factor4.Level1	0.216
	(0.081)
Factor5.Level1	0.145
	(0.082)
Factor5.Level2	0.112
	(0.086)
Factor6.Level1	0.122
	(0.101)
Factor6.Level2	0.192
	(0.091)
Factor6.Level3	0.067
	(0.113)
Factor6.Level4	0.107
	(0.100)
Factor6.Level5	0.151
	(0.089)
Factor6.Level6	0.106
	(0.111)
Factor6.Level7	0.088
	(0.115)
Factor6.Level8	−0.132
	(0.162)
Factor6.Level9	0.179
	(0.100)
Factor7.Level1	0.121
	(0.086)