Batch Adaptive Designs to Improve Efficiency in Social Science Experiments*

Matthew Blackwell† Nicole E. Pashley‡ Dominic Valentino§

December 7, 2023

Abstract

Experiments are vital for assessing causal effects, but their high cost often leads to small, suboptimal sample sizes. We show how a particular experimental design—the Neyman allocation—can lead to more efficient experiments, achieving similar levels of statistical power as traditional designs with significantly fewer units. This design relies on unknown variances, and so previous work has proposed what we call the *batch adaptive Neyman allocation* (BANA) design that uses an initial pilot study to approximate the optimal Neyman allocation for a second larger batch. We extend BANA to multiarm experiments common in political science, derive an unbiased estimator for the design, and show how to perform inference in this setting. Simulations verify that the design's advantages are most apparent when the outcome variance differs by treatment conditions. Finally, we review the heteroskedasticity of recent experimental studies and find that political scientists using BANA could achieve sample size savings of 15–30%.

*Working paper, comments welcome. Thanks to Alexander Coppock, Drew Dimmery, Gary King, Brendan Nyhan, and Maya Sen for helpful suggestions and feedback. Any errors remain our own.

†Department of Government and Institute for Quantitative Social Science, Harvard University. web: http://www.mattblackwell.org email: mblackwell@gov.harvard.edu

‡Department of Statistics, Rutgers University. email: nicole.pashley@rutgers.edu

§Department of Government, Harvard University. email: dvalentino@g.harvard.edu

1 Introduction

Experimental studies are a valuable, if costly, tool for researchers in the social sciences. The sometimes enormous costs of recruiting participants, implementing the treatments, and measuring outcomes often force researchers to compromise statistical power by keeping their sample sizes small. Researchers can improve the efficiency of their statistical estimates in the design of their experiment by using tools such as blocking (Fisher, 1935; Imai, King and Stuart, 2008; Pashley and Miratrix, 2022), but these tools are often infeasible because they require covariate measurement before randomization. Within-subjects designs that measure the outcome pre- and post-treatment can also make experiments more powerful but require panel data on respondents and can suffer from differential attrition.

We focus on an overlooked aspect of experimental design that can increase statistical efficiency: the relative allocation of treatment. The literature on experimental design in statistics has long known that researchers can achieve higher statistical power by implementing an optimal design known as the Neyman allocation that assigns more units to treatment arms that have outcomes with higher variances. Unfortunately, this optimal design is infeasible because it depends on unknown outcome variances; to address this, scholars have proposed using an initial pilot batch to estimate the optimal Neyman allocation in future batches (Robbins, 1952; Hahn, Hirano and Karlan, 2011). This paper studies the properties of this design, which we call the *batch adaptive Neyman allocation* (BANA) design, and shows how researchers can leverage the Neyman allocation to make their experiments more efficient, sometimes vastly so.

Our paper makes five contributions to the study of adaptive experimental designs in political science. First, we derive the finite-sample relative efficiency of the Neyman allocation compared to the traditional uniform allocation, allowing researchers to understand when the Neyman allocation is likely to improve their designs. In short, the Neyman allocation will be more efficient when the outcome is more heteroskedastic across treatment conditions. Second, we extend the BANA de-

sign in the potential outcome framework beyond the binary treatment setting to allow for multiarm studies common in the social sciences. Third, we find that BANA performs poorly with rare binary outcomes and develop a shrinkage estimator for the optimal allocation weights that vastly improves finite-sample performance. Fourth, we show how to perform inference for this design, using a standard stratified difference-in-means estimator that is unbiased for the average treatment effect in spite of the adaptive design. Our inference methods are valid in finite samples and do not rely on asymptotic results, which may not be reasonable for many experiments. Finally, through simulations and a literature review, we show that researchers in the political science could gain up to 15–30% improvements in relatively efficiency in their main batches using the BANA design.

Adaptive experimental designs have been widely used throughout the sciences and are especially popular in commercial contexts (for a review focused on political science, see Offer-Westort, Coppock and Green, 2021). The BANA design differs from common adaptive designs like the multi-armed bandit (Berry and Fristedt, 1985) in optimizing the efficiency of the experiment rather than attempting to maximize the average outcome. The latter is often very useful for commercial or medical settings where researchers might be interested in which ad causes the highest engagement or which drug produces the best patient outcome. But finding the optimal treatment arm is often less relevant in the context of social science research, where the precise estimation of causal contrasts is often more important than identifying the best performing arm. Additionally, we recommend a small number of batches—usually just two—to make the implementation of the design much simpler than real-time adaptive designs that update the treatment allocation for each new unit. This allows a broader application of the design with standard commercial survey administration platforms. Furthermore, researchers can leverage the small pilot studies they are already running to implement the BANA design, making its incorporation into experimental design straightforward. Another difference with the bandit literature is that, as we show below, the batching nature of the design means that we can use standard estimators for stratified randomized experiments viewing the batches as strata.

The paper proceeds as follows. Section 2 reviews the literature on adaptive experimental designs.

In Section 3, we describe the basic setting and derive the relative efficiency of the Neyman allocation compare to the uniform. In Section 4, we introduce the BANA estimator, derive its statistical properties, and develop a modified version of the approach with binary outcomes. We present simulation evidence for the usefulness of the design in Section 5. Section 6 presents a meta-analysis of several recent experimental papers in political science and how much a BANA design, if conducted, might have improved the efficiency of the experiment. Finally, in Section 7, we provide practical advice on using adaptive designs for social scientists, and in Section 8, we end with directions for future research.

2 A Review of Adaptive Experimental Designs

Neyman (1934) was the first to observe optimal sampling designs, focusing on randomly sampling from a population of strata, where the goal is to estimate the mean of a variable in the population. Neyman found that the optimal proportion to allocate to a particular stratum was proportional to the standard deviation of the variable of interest in that stratum. Soon after, Sukhatme (1935) evaluated a scheme where a small sample was taken first and then used to estimate the population variances and thus the optimal design. Sukhatme (1935) approximated the distribution of the estimated variances under a normality assumption to determine how likely this pilot study approach would be to outperform a uniform allocation. Solomon and Zacks (1970) provides a comprehensive overview of the literature on optimal sampling designs from both the finite-sample and Bayesian perspectives. More recently, Melfi, Page and Geraldes (2001) and Hu and Zhang (2004) proposed analyses of two-arm and multi-arm adaptive designs that are continuously updated throughout the experiment rather than batched like ours, and Rosenman and Owen (2021) shows how one might use an observational study to inform the optimal Neyman allocation in a follow-up stratified randomized experiment. Dimmery (2019) proposes a similar design to ours that allows researchers to avoid explicitly incorporating the design into the analysis, but requires several additional assumptions that we side-step by explicitly adjusting for the design through stratification.

A more recent literature has investigated how two-batch designs can be used to find optimal stratification based on covariates (Cytrynbaum, 2021; Tabord-Meehan, 2021; Armstrong, 2022). These techniques are ideal when covariate information will be available at the time of randomization, but this often is not the case for political science experiments. Tabord-Meehan (2021) proposes to stratify (that is, partition) the (continuous) covariate space and selects the treatment allocations within those strata that will optimize the variance of an inverse-probability weighting estimator of the ATE. The optimal stratification tree is found using an optimization routine on the first-batch data. Cytrynbaum (2021) proposes a similar strategy that first creates "local groups" based on a discretized version of the estimated optimal propensity score and then formulated matched strata within these local groups based on covariates. Randomization then occurs within these strata.

Inference in adaptive designs has been a challenge because adaptivity creates dependence across outcomes and propensity scores moving arbitrarily close to zero can lead to non-normal asymptotic distributions. Some authors have set asymptotic normality aside in favor of finite-sample bounds using martingale concentration inequalities (Howard et al., 2021) while others have modified augmented inverse probability weighting estimators to stabilize the asymptotic distribution and regain normality (Hadad et al., 2021). Our clipping of the propensity scores in the second batch helps to avoid these issues at the expense of moving away from the optimal design when the heteroskedasticity is extreme.

Zhang, Janson and Murphy (2020) develop a batched ordinary least squares approach, very similar to our proposed estimator and shows that when the estimator is stratified by batch, the asymptotic distribution of the treatment effect estimator is consistent and asymptotically normal. Their setting is slightly different from ours in that they focus on the case where the sample sizes are constant across batches and Bernoulli randomization is used to assign treatment within batches, whereas we focus on a small first batch and complete randomization within batches. In addition, they make a conditional homoskedasticity assumption that makes it difficult to apply to our setting.

3 Setting and Background

We consider an experiment on units $i \in \{1, \ldots, N\}$. Let $D_i \in \{0, 1, \ldots, J\}$ be the treatment assigned to unit i and $Y_i(d)$ be the potential outcome for unit i when $D_i = d$. We take a finite-sample perspective and consider these potential outcomes as fixed and we make the usual SUTVA assumption (Rubin, 1980) that the observed outcome $Y_i = Y_i(D_i)$. We define the sample average and variances of these potential outcomes as

$$\overline{Y}(d) = \frac{1}{N}\sum_{i=1}^{N} Y_i(d) \qquad S^2(d) = \frac{1}{N-1}\sum_{i=1}^{N}\left\{Y_i(d) - \overline{Y}(d)\right\}^2,$$

where the targets of inference are the sample average treatment effects, $\tau(d, d') = \overline{Y}(d) - \overline{Y}(d')$. Once treatment has been assigned, we can calculate the sample outcome means in each arm as

$$\overline{Y}^{\text{obs}}(d) = N_d^{-1}\sum_{i=1}^{N} \mathbb{I}(D_i = d)Y_i,$$

where N_d is the number of units in arm d. The difference in means estimators for the sample average treatment effects are $\widehat{\tau}(d, d') = \overline{Y}^{\text{obs}}(d) - \overline{Y}^{\text{obs}}(d')$.

To highlight how the experimental design can impact the efficiency of an experiment, we explore two different randomization allocations for a single batch. First, the *uniform allocation* assigns the same number of units, $N_d = N/(J+1)$ units, to each of the treatment arms so that

$$\pi_u(d) = \mathbb{P}(D_i = d) = \frac{1}{J+1}.$$

The second strategy is the *Neyman allocation* (Neyman, 1934; Cochran, 1977) which allocates in proportion to the standard deviation of the potential outcomes under each arm,

$$\pi_{na}(d) = \mathbb{P}(D_i = d) = \frac{S(d)}{\sum_{j=0}^{J} S(j)}.$$

This Neyman allocation minimizes the average of the variances of the sample outcome means in each treatment arm, $(J+1)^{-1}\sum_{j=0}^{J}\mathbb{V}[\overline{Y}^{\text{obs}}(d)]$, which also minimizes the average variance across all possible treatment effect contrasts $\widehat{\tau}(d, d')$.

We can compare the variance of the difference in means estimator under the two proposed allocations, letting $\mathbb{V}_u[\cdot]$ be the finite-sample variance under the uniform allocation and $\mathbb{V}_{na}[\cdot]$ be the finite-sample variance under the Neyman allocation. For simplicity of exposition, we focus on the binary treatment case under complete randomization where we can assign the exact number of optimal units. Letting $\widehat{\tau}_1 = \widehat{\tau}(1,0)$, the variances in this setting are

$$\mathbb{V}_u[\widehat{\tau}_1] = \frac{2S^2(1) + 2S^2(0) - S^2(1,0)}{N} \qquad \mathbb{V}_{na}[\widehat{\tau}_1] = \frac{\{S(1) + S(0)\}^2 - S^2(1,0)}{N},$$

where $S^2(1,0) = (N-1)^{-1} \sum_{i=1}^{N} \left\{ Y_i(1) - Y_i(0) - (\overline{Y}(1) - \overline{Y}(0)) \right\}^2$ is the variance in the treatment effect across units (see, e.g., Imbens and Rubin, 2015). It is easy to show that $\mathbb{V}_u[\widehat{\tau}_1] - \mathbb{V}_{na}[\widehat{\tau}_1] = N^{-1}(S(1) - S(0))^2 \geq 0$ with equality only holding when $S(1) = S(0)$ which is exactly when the Neyman allocation and the uniform allocation are the same. Thus, the Neyman allocation will lead to a more efficient estimator of the sample average treatment effect unless there is homoskedasticity across treatment arms.

How much more efficient is the Neyman allocation? We can answer this with the relative efficiency of the Neyman allocation, $\mathbb{V}_{na}[\widehat{\tau}_1]/\mathbb{V}_u[\widehat{\tau}_1]$, which is a scale-free measure of how many fewer units the Neyman allocation requires compared to the uniform allocation to achieve the same variance. This relative efficiency depends on both the degree of heteroskedasticity, as measured by the treatment-control ratio of standard deviations, $\delta = S(1)/S(0)$, and the (unidentified) correlation between the potential outcomes, ρ. We show in the Supplemental Materials that we can write the relative efficiency as

$$\frac{\mathbb{V}_{na}[\widehat{\tau}_1]}{\mathbb{V}_u[\widehat{\tau}_1]} = \frac{2\delta(1+\rho)}{1 + \delta^2 + 2\delta\rho}. \tag{1}$$

As heteroskedasticity across treatment conditions increases (δ moves away from 1), the Neyman allocation becomes more efficient than the uniform. For example, if the variance in the treated group is twice that of the control group and the potential outcomes are uncorrelated, we could use 20% fewer observations in a Neyman allocation compared to a uniform allocation to achieve the same variance. If the potential outcomes are correlated, this comparison will change but even with a correlation of

0.5, the Neyman allocation would require 14% fewer observations. Generally, we should expect variances will differ between treatment arms when treatment effects differ across units, and thus that is when we should expect Neyman allocation to be advantageous.

Of course, the optimality of the Neyman allocation requires knowledge of the true variance of the potential outcomes in each arm, so below we consider a batch adaptive design that allows us to leverage the Neyman allocation without sacrificing sample size.

3.1 Control-augmented Neyman Allocation

The above derivation of the Neyman allocation assumes that we care about all treatment arm comparisons equally, including those between active treatments. When our interest is focused more on minimizing the variance of treatment effect (compared to the control condition, $D_i = 0$), then we need a slightly different allocation. In that case the allocation that minimizes the average variance weights more heavily the control group to acknowledge its role in all of the effect comparisons. Following Offer-Westort, Coppock and Green (2021), we call this a control-augmented Neyman allocation and its allocation is

$$\pi_{ca}(0) = \frac{\sqrt{J}S(0)}{\sqrt{J}S(0) + \sum_{j=1}^{J} S(j)}, \qquad \pi_{ca}(d) = \frac{S(j)}{\sqrt{J}S(0) + \sum_{j=1}^{J} S(j)} \text{ for } d \in \{1, \ldots, J\}.$$

We show the derivation of this result in the Supplemental Materials. This allocation highlights how the uniform allocation may be inefficient in multi-arm studies where the objects of inference are all comparisons to a control group. In those cases, even if there is no heteroskedasticity across treatment conditions, there is still an efficiency advantage to allocating more units to control than to the other conditions. The above result implies that even with just two treatment arms and a control arm, if the variances are all (roughly) equal, the control arm should have roughly 40% more units than either of the other two arms to reach maximum efficiency.

Of course, it is not always clear that comparisons to control are the only object of interest in multi-arm studies. Scholars may use such comparisons as a basis for comparing different treatments,

but will often go on to compare different treatment arms directly to see if one has a statistically significantly higher outcome than another. In those cases, the Neyman allocation may still be desirable.

4 The BANA Design

We now describe the BANA design, first proposed by Robbins (1952) though that discussion is rather informal. Hahn, Hirano and Karlan (2011) formalized the design for the binary treatment case under a superpopulation approach and we extend this to multi-arm experiments and focus on a finite-population setting. The BANA design splits our units into two batches $\{1,\ldots,N_1\}$ and $\{N_1{+}1,\ldots,N\}$, where $N_2 = N - N_1$. In the first batch, we perform a completely randomized design with uniform allocation so that $N_{1d} = N_1/(J+1)$ units are allocated to treatment arm d. From this arm, we estimate the variances of the outcome in each arm as

$$s_1^2(d) = \frac{1}{N_{1d}-1}\sum_{i=1}^{N_1}\mathbb{I}(D_i = d)\left(Y_i - \overline{Y}_1^{\text{obs}}(d)\right)^2,$$

where $\overline{Y}_1^{\text{obs}}(d)$ is the average outcome in the $D_i = d$ group in the first batch. We use these estimated variances to estimate the Neyman allocation proportions,

$$\widehat{\pi}_{na}(d) = \frac{s_1(d)}{\sum_{j=0}^{J} s_1(j)},$$

which we then use for the treatment assignment in the second batch. With a Bernoulli randomized design, we can simply draw D_i as an i.i.d. categorical variable with probability vector $(\widehat{\pi}_{na}(0),\ldots,\widehat{\pi}_{na}(J))$. With a completely randomized design we have to choose the (integer) allocation of units that best approximates those proportions:

$$\underset{(n_{20},\ldots,n_{2J})}{\arg\min} \quad \sum_{j=0}^{J}\left|n_{2j} - N_2\widehat{\pi}_{na}(j)\right|, \text{ such that } n_{2j}\in\{Q,\ldots,N_2 - Q\times J\}, \sum_{j=0}^{J} n_{2j} = N_2, \qquad (2)$$

where we use the lower case for n_{2j} because these allocation numbers are functions of the treatment assignment in the first batch and are therefore random, even in the finite sample setting. Here, we have

allowed for a minimum number of units that can be allocated to a given arm, Q.[1] This can be helpful in the extremely rare cases where there is extreme in-sample heteroskedasticity across treatment arms in the first batch that would imply that very few (or no) observations be assigned to one of the arms. Several fast algorithms exist for solving this optimization routine exist such as Friedrich et al. (2015) (see Ravichandran et al., 2023, who use these integer optimizations in this setting), but it is also possible to use a crude rounding approximation. Once we complete the second batch, we observe the outcomes as

$$\overline{Y}_2^{\text{obs}}(d) = \frac{1}{n_{2d}} \sum_{i=N_1+1}^{N} \mathbb{I}(D_i = d) Y_i.$$

The adaptive nature of the BANA designs creates dependence between units, complicating the analysis of the experiment. Naively combining both batches and analyzing them as a single experiment is problematic because treatment assignment for some units (those in the second batch) depends on the realized outcomes of other units (those in the first batch). Alternatively, we could drop the first batch entirely and only analyze the second batch conditional on the first batch, but this sacrifices statistical power.

A better way to analyze the BANA design without bias while maximizing statistical power is to treat it as a stratified randomized design where the batch is the stratifying variable (Zhang, Janson and Murphy, 2020). Let $\widehat{\tau}_b(d, d') = \overline{Y}_b^{\text{obs}}(d) - \overline{Y}_b^{\text{obs}}(d')$ be the estimated treatment effect in batch b and define the usual stratified estimator as

$$\widehat{\tau}_s(d, d') = \left(\frac{N_1}{N_1 + N_2}\right) \widehat{\tau}_1(d, d') + \left(\frac{N_2}{N_1 + N_2}\right) \widehat{\tau}_2(d, d'),$$

with variance estimator

$$\widehat{\mathbb{V}}[\widehat{\tau}_s(d, d')] = \left(\frac{N_1}{N_1 + N_2}\right)^2 \widehat{\mathbb{V}}_1[\widehat{\tau}_1(d, d')] + \left(\frac{N_2}{N_1 + N_2}\right)^2 \widehat{\mathbb{V}}_2[\widehat{\tau}_2(d, d')],$$

where $\widehat{\mathbb{V}}_b[\cdot]$ are the usual conservative Neyman variance estimators within each batch:

$$\widehat{\mathbb{V}}_1[\widehat{\tau}_1(d, d')] = \frac{s_1^2(d)}{N_{1d}} + \frac{s_1^2(d')}{N_{1d'}}, \qquad \widehat{\mathbb{V}}_2[\widehat{\tau}_2(d, d')] = \frac{s_2^2(d)}{n_{2d}} + \frac{s_2^2(d')}{n_{2d'}}.$$

[1] While we focus on a completely randomized design, these ideas can be straightforwardly generalized to a stratified randomized design.

In the Supplemental Materials we show that this variance estimator is unbiased when treatment effects are constant and conservative for the true variance otherwise.[2]

Finally, we note that while we focus on the case of just two batches, it would be easy to extend this to larger numbers of batches that allow better estimation of the optimal allocation. Our simulation results below, however, indicate that we can obtain much of the benefit of the optimal design with a relatively small first batch, calling into question the need for further batches.

4.1 BANA with Binary Outcomes

Using the BANA design with a binary outcome variable presents a unique challenge due to the dependence between the mean and variance of Bernoulli random variables. In particular, when initial batch sizes are small and probabilities of success, $\mathbb{P}(Y_i = 1)$, are near zero or one, large finite-sample errors in estimating the means in each condition can lead to highly variable estimates of the weights for the second batch. These unstable estimates of the weights can produce second-batch treatment allocations that perform worse than the uniform allocation.[3]

To address this issue in the binary outcome case, we adopt an empirical Bayesian approach that flexibly shrinks the weights toward the uniform allocation. Typically, we would estimate the standard deviations of each arm with the usual sample standard deviation estimator. We instead adjust this estimator by shrinking toward the grand outcome mean across all conditions, $\widehat{p}_{\text{grand}} = N_1^{-1} \sum_{i=1}^{N_1} Y_i$. In particular, we estimate the standard deviation using proportions that are shrunk toward the grand mean,

$$s_{1\lambda}(d) = \sqrt{\frac{N_{1d}}{N_{1d} - 1} \widehat{p}_\lambda(d)(1 - \widehat{p}_\lambda(d))},$$

where

$$\widehat{p}_\lambda(d) = \left(\frac{\lambda}{N_{1d} + \lambda}\right) \widehat{p}_{\text{grand}} + \left(\frac{N_{1d}}{N_{1d} + \lambda}\right) \overline{Y}_1^{\text{obs}}(d),$$

[2] With a completely randomized experiment and a binary treatment, our stratification estimator is equivalent to the weighting estimator of Hahn, Hirano and Karlan (2011).

[3] See the simulation section for more on this problem.

where the λ parameter captures how much we shrink toward the grand mean. This shrinkage proportion estimator $p_\lambda(d)$ adds λ artificial observations to each treatment arm that have outcomes exactly at the overall mean, $\widehat{p}_{\text{grand}}$. The higher the value of λ, the closer the shrinkage estimator will be to the grand mean. We can then plug these estimates of the standard deviations into the formula for the Neyman weights $\widehat{\pi}_{na}(d)$ given above.[4]

4.2 Randomization Inference for the BANA Design

One can use the variance estimator above to construct confidence intervals using the usual large-sample normal approximation (which works well in our simulations), but we can also leverage the design to implement a randomization inference approach that is valid for all sample sizes. Focusing on the binary treatment setting, we now describe a method for inference on the BANA design that leverages randomization inference under a sharp null hypothesis,

$$H_0 : Y_i(1) = Y_i(0) \quad \forall i,$$

which can be easily generalized to the multi-arm setting. Under this sharp null, the treatment has no effect for any unit and the optimal design is thus the uniform allocation.

There are two possible ways to conduct randomization inference in this setting. The simplest is to condition on the first-batch data and perform a standard permutation test on the second batch which obviously will have lower power due to the omission of the first-batch data. The second approach is to replicate both stages of the randomization procedure repeatedly. Specifically, repeat the following steps for $r = 1, \ldots, R$:

1. Use complete randomization to assign N_{11} units from the first batch to treatment ($\widetilde{D}_{i,r} = 1$) and N_{10} to control ($\widetilde{D}_{i,r} = 0$).

[4]This shrinkage estimator for the group means that we plug into the weights is also an empirical Bayes estimator. In particular, we model the data, $\sum_{i:D_i=d, i<N_1} Y_i$, as distributed Binomial(N_{1d}, p_d) and place a prior distribution over the success probability as $p_d \sim \text{Beta}(\alpha, \beta)$, where $\alpha = \widehat{p}_{\text{grand}}\lambda$ and $\beta = (1 - \widehat{p}_{\text{grand}})\lambda$.

2. Calculate the observed mean and standard deviation of the treated and control group

$$\widetilde{Y}_{1,r}^{\text{obs}}(d) = \frac{1}{N_{1d}} \sum_{i=1}^{N_1} \mathbb{I}(\widetilde{D}_{i,r} = d) Y_i, \qquad \widetilde{s}_{1,r}^2(d) = \frac{1}{N_{1d} - 1} \sum_{i=1}^{N_1} \mathbb{I}(\widetilde{D}_{i,r} = d) \left(Y_i - \widetilde{Y}_{1,r}^{\text{obs}}(d) \right)^2.$$

3. Calculate the estimated optimal allocation $\widetilde{\pi}_{na,r}(d) = \widetilde{s}_{1,r}(d)/(\widetilde{s}_{1,r}(1) + \widetilde{s}_{1,r}(0))$ and conduct a complete randomization assigning $\widetilde{n}_{21,r} = \lfloor N_2 \widetilde{\pi}_{na,r}(d) \rfloor$ units from the second batch to treatment and $\widetilde{n}_{20,r} = N_2 - \widetilde{n}_{21,r}$ to control.

4. Calculate the observed mean of the second batch, $\widetilde{Y}_{2,r}^{\text{obs}}(d)$ and estimated treatment effect

$$\widetilde{\tau}_r = \left(\frac{N_1}{N} \right) \left\{ \widetilde{Y}_{1,r}^{\text{obs}}(1) - \widetilde{Y}_{1,r}^{\text{obs}}(0) \right\} + \left(\frac{N_2}{N} \right) \left\{ \widetilde{Y}_{2,r}^{\text{obs}}(1) - \widetilde{Y}_{2,r}^{\text{obs}}(0) \right\}.$$

5. Use the estimated treatment effect to calculate a test statistic, $\widetilde{T}_r$.

We can then compare the observed value of the test statistic in the data, T, with the randomization distribution of the test statistic $\widetilde{T}_r$. We can calculate a p-value for the sharp null hypothesis in the usual way by $\frac{1}{R} \sum_{r=1}^{R} \mathbb{I}(T > \widetilde{T}_r)$.

It is also possible to formulate confidence intervals for the estimated effect by considering a grid of sharp null hypotheses all assuming a constant treatment effect. Under each of these null hypotheses, we can determine all of the missing potential outcomes and we simply modify the above algorithm to use these imputed potential outcomes rather than the observed outcomes. Then, we can find the set of null hypotheses that cannot be rejected at level α to form a $(1 - \alpha) \times 100$ percent confidence interval.[5]

Finally, in the above procedure, we fix the units that belong to the first and second batch, rather than allowing, for example, unit 1 to be potentially assigned to the first or second batch. This is likely consistent with most empirical designs where the researcher has control over what treatment assignment each unit receives but not who is in the first or second batch. If the units across batches are exchangeable (possibly because the batching was randomly assigned), one could modify the above procedure to allow for units to be assigned either to the first or second batch.

[5]These types of randomization-inference-based intervals are sometimes referred to as Fisher intervals.

5 Simulation Studies

We now present a pair of simulation studies that illustrate the advantages of the BANA design with a binary treatment. We simulate experiments with Gaussian outcomes and binary outcomes separately to compare the performance of the BANA design in both the continuous and discrete outcome cases.

Regardless of the type of outcome, we simulate three different treatment allocation schemes. The first is a simple uniform allocation of all respondents equally to treatment and control. The second is the BANA design described above. The third allocation scheme is the BANA design with oracle Neyman allocation, using the true standard deviation of the outcome in the treatment and control conditions. Note that the oracle Neyman allocation is infeasible but the theoretical best. The BANA design is our best approximation of this infeasible oracle. Comparing BANA to the oracle allows us to observe how much noise the estimation of variance introduces into estimates of the ATE. Comparing the oracle to the uniform allows us to observe how much potential gains we can get if we use adaptive designs. All three allocation schemes use complete randomization within batches. We set the minimum number of units that can be allocated to a given treatment arm to $Q = 2$, allowing us to estimate variance within each experimental arm.

5.1 Gaussian Outcome

The data generating process (DGP) for the simulations with a Gaussian outcome, regardless of type of treatment assignment, draws the potential outcomes as

$$Y_i(0) \sim \mathcal{N}(0, \sigma_c^2), \qquad Y_i(1) \sim \mathcal{N}(0.1, \sigma_t^2),$$

for all $i \in \{1, \ldots, N\}$. We define the observed outcome as $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$ where D_i indicates treatment for unit i. This DGP has an ATE of 0.1 but individual-level treatment effect heterogeneity.

For each allocation scheme, we vary the ratio of σ_t to σ_c from 0.05 to 1, which controls the heteroskedasticity across treatment conditions. At ratios closer to zero, the outcome under treatment

is much less variable than under control, and at ratios closer to one the variance of the outcome becomes more similar. The performance of the BANA design does not depend on which treatment arm has the higher variance (since the Neyman allocation treats all arms equivalently) so we focus on the values of the ratio between zero and one (i.e., higher variance under control). Importantly, while we vary the relative variability of the outcomes in each of the conditions, we hold the overall variability of the outcomes across conditions constant at $\sigma = 1$. We also vary the initial batch size (N_1) over the values $\{10, 25, 50\}$ to investigate how estimation uncertainty from the initial batch could affect our inferences, though we present a wider variety of batch and sample sizes in the Supplemental Materials. For each combination of parameters, we draw $R = 30,000$ iterations, redrawing the potential outcomes and assigning treatment in each replication. Using the simulated data, we estimate the ATE and its standard error with the stratified estimators defined above. These simulations are based on a superpopulation approach, which we can view as approximating the average performance of the estimators across finite samples. In particular, in each simulation new potential outcomes are drawn and treatment is reassigned. Root mean squared error (RMSE) and coverage are taken with respect to the "superpopulation" parameters that generate potential outcomes.

Our main comparison between allocation schemes is made in terms of power to detect the true ATE. Specifically, since power is $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ is true})$ and we have constructed H_1 to be true, we simply calculate the proportion of estimates for which $p < .05$ using a t-test under the large-sample normal approximation. We also calculate root mean squared error (RMSE) and 95% confidence interval coverage probabilities under each allocation scheme.

Figure 1 presents the results of this simulation. Beginning with statistical power in panel a), we see that the BANA design has higher power than a uniform design when the variance of the outcome is very different across conditions. When using a very small initial batch ($N_1 = 10$), BANA underperforms both the oracle counterpart and the uniform allocation because estimation error leads to inefficient allocation of units. However, we only need a slightly larger initial batch to combat this problem—when using a still very modest initial batch size of $N_1 = 25$, BANA does about at least as

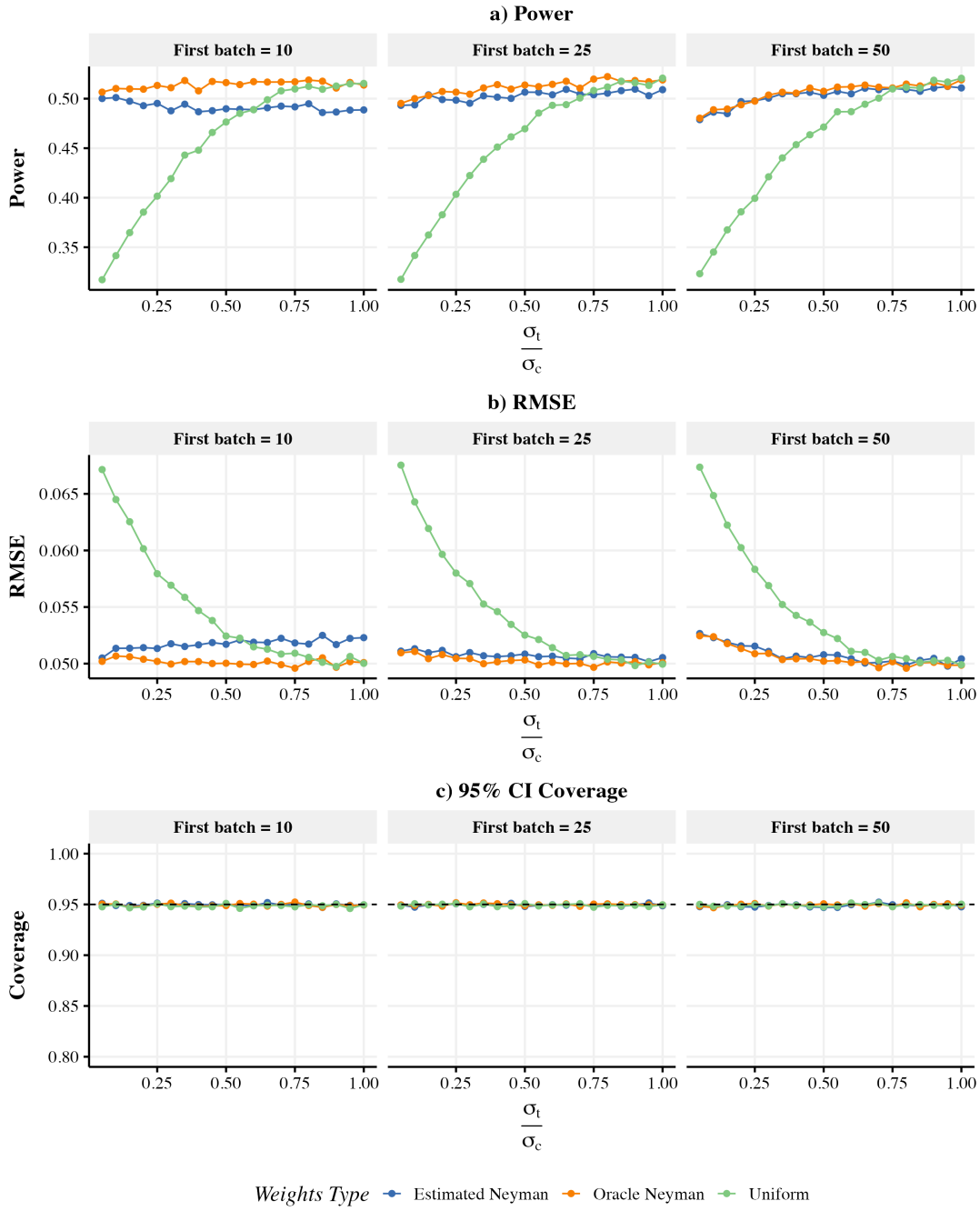

Figure 1: Gaussian outcome simulation results. N = 400, R = 30,000. Horizontal panels differ by the size of the first batch. The x-axis displays the ratio of the standard deviations of the outcome in the treatment and control conditions, but note that overall variability of the outcome is constant. Colors indicate allocation scheme. The outcomes are a) power to detect the treatment effect, b) root mean squared error, and c) 95% CI coverage probability.

well as both the oracle and uniform designs at any level of similarity between the variances. With an even larger initial batch size ($N_1 = 50$) this pattern continues, but power dips down at lower levels of similarity. This is caused by the fact that an increasing proportion of the units used in estimating the ATE are being allocated inefficiently in the initial batch, leading the power curves to be more similar across allocation types.

A similar pattern can be observed in Figure 1(b) with RMSE. Again, we see BANA performing at least as well as the uniform design except for at larger ratios in the smallest initial batch size ($N_1 = 10$). Initial batch sizes of 25 and 50 produce very similar results. Lastly, panel c) shows that all the allocation strategies have roughly nominal coverage of the 95% confidence intervals.

Based on these results, we recommend an initial batch size of at least 25–50 units. For studies with larger total sample sizes, the researcher may be willing to allocate more units to the first stage to aid in estimation of the Neyman probabilities, but very little efficiency is gained by using more than 50 units in the initial batch. Figure SM.7 in the Supplemental Materials shows simulation results on a wider array of first and second batch sizes and Figure SM.8 shows the same results but using only the second batch for estimation, both of which suggest similar conclusions. We note, however, that the simulations are not exhaustive and different data generating mechanisms may lead to different batch size recommendations. Furthermore, our results show that a uniform allocation in the second batch is a reasonable approach when the heteroskedasticity is not too severe.

5.2 Binary Outcome

In the binary simulation case we use the following DGP:

$$Y_i(0) \sim \text{Bern}(p_0), \qquad Y_i(1) \sim \text{Bern}(p_0 + 0.1),$$

for all $i \in \{1, \ldots, N\}$ where the observed outcome is defined as above. We control average probabilities of success in each experimental condition, and these also determine variability of the outcome. Specifically, we vary p_0 from 0.05 to 0.4 with a constant treatment effect of 0.1, meaning the probability of success under treatment varies from 0.15 to 0.5. Since the variance of a Bernoulli random

variable is maximized at 0.5 and the BANA design is agnostic to which condition is more variable, we can safely ignore success probabilities between 0.5 and 1 in these simulations. Because the outcome variance is tied to the probability of success, the overall variability cannot be held constant if the effect size is also held constant, in contrast to the Gaussian simulations.

Importantly, as described in Section 4.1, we add a number of artificial observations equal to the grand mean of the outcome to the treatment and control conditions in the first batch. This guards against inefficient allocations caused by estimation error when the initial batch size is small or the probability of success is extreme.

Additionally, we impose a minimum probability of allocation to treatment in the second batch of 0.1, ensuring that each condition is at least 10% of the overall batch size. This guards against cases in which the variance of the outcome in one of the conditions is extremely small or zero, which occurs regularly in small initial batch sizes with rare outcomes. This step is taken directly after estimating the weights with artificial observations.

As in the Gaussian simulations, we vary the initial batch size, this time over the values $\{50, 100, 200\}$. Discussed in more detail later, the BANA design needs larger initial batch sizes to accurately estimate variances in the binary case. For each combination of parameters, we draw $R = 30,000$ iterations, again redrawing the potential outcomes and assigning treatment in each replication. We estimate the ATE and its standard error in exactly the same way as above.

Simulation results with a binary outcome are shown in Figure 2. We find much more modest gains under these conditions, sometimes even resulting in allocations that are less efficient relative to uniform when batch size and the probability of success are low.[6] This likely results because heteroskedasticity across experimental conditions is constrained by the fact that the outcome is binary combined with difficulty estimating probabilities of success in low first-batch sizes with rare outcomes. For context, the minimum standard deviation ratio is .05 in the Gaussian case, while the

[6]Note that this problem is significantly attenuated by the addition of observations equal to the grand mean before calculating treatment allocation weights for the second batch. See Figures SM.5 and SM.6 to compare. While main results in the Gaussian case are presented in terms of power, we use relative efficiency in the binary case to better visualize these modest gains relative to simulation noise.

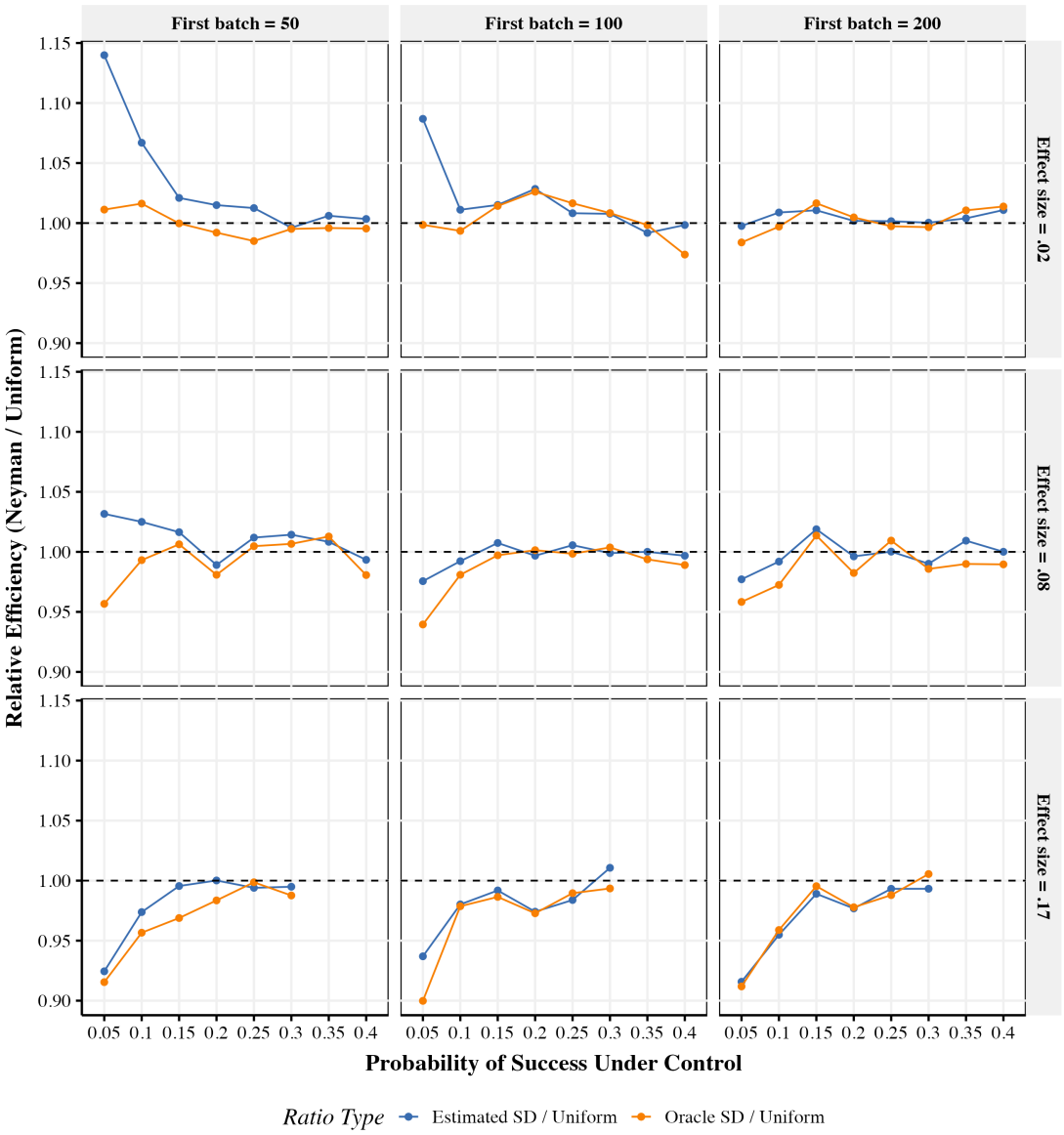

Figure 2: Binary outcome simulation results. $N = 1,000, R = 30,000$. Horizontal panels differ by the size of the first batch and vertical panels differ by the effect size, informed by the meta-analysis below. The x-axis displays the probability of success in the control condition; note that unlike the Gaussian case, total variability is not constant. The y-axis displays the ratio of the standard error of the treatment effect in the Neyman case (either oracle or estimated) to the uniform case. Colors indicate allocation scheme relative to uniform. Minimum allocation to each condition in the second batch is set to .1 and five additional observations were added to each condition in the first batch.

minimum ratio in the binary case is about .37. Still, we do find that experiments with large effect sizes can save between 5 and 8% of their sample size using the BANA design.

6 Meta-analysis

Simulations provide evidence that the BANA design can improve efficiency, but how do these simulation compare with real studies? To gauge the potential for efficiency gains from the BANA method in real studies, we collected or computed the variance of the outcome in each treatment arm in several experimental studies published in top journals in political science. The studies, listed in Table 1, span all substantive fields of political science and include both survey and field experiments. Table 1 shows the control standard deviation of the outcome and the standard deviation of the outcome in the most variable treatment arm. We use the relative efficiency formula in (1) to compute the proportional reduction in variation that would be obtained under a BANA design vs uniform allocation in each study, assuming uncorrelated potential outcomes. Larger values indicate more savings under BANA. This ratio also roughly approximates the reduction in sample size under the BANA design to achieve the same precision on the estimate of the treatment effect, though we note that these reductions only apply to the second batch. Further, these reductions are based on treating estimated variance as the truth, and thus reflect gains in an oracle BANA design. We expect in practice to have somewhat smaller gains due to the estimation of variance.

Importantly, we assume that every observation in each study could be allocated according to the BANA design. Because our recommendation of initial batch size varies according to the type of study design, we calculate expected savings as if each study listed had been previously informed by its own pilot study from which outcome variances in each condition were estimated.

Results of the meta-analysis are shown in Table 1. We find that in most of the political science experiments under consideration, treatment has only a small effect on the variance of the outcome, leaving the BANA design with little leverage to improve upon the standard design. However, we also find that improvements to precision from the BANA design can be considerable—researchers can reduce the variance of the treatment effect estimate by 30% in some cases, leading to a comparable reduction in the sample size needed to achieve a given amount of statistical power as well. These

	Study	Control SD	Treatment SD	Variance Reduction (%)	Type of Study
1	Siegel and Badaan (2020)	5.643	27.212	30.10	Six arm experiment
2	Vernby (2019)*	0.152	0.398	16.70	2x8 factorial
3	Eble et al. (2021)	14.2	22.3	4.70	Two arm experiment
4	DeVreese (2004)	0.88	0.6	3.50	Two arm experiment
5	Broockman (2013)*	0.5	0.359	2.60	Two arm experiment
6	Eggers et al. (2017)*	0.494	0.372	1.90	2x2 factorial
7	Goff et al. (2017)	3.34	2.57	1.70	Two arm experiment
8	Holman et al. (2016)	16.522	21.019	1.40	2x4 factorial
9	Simas and Murdoch (2020)	0.946	1.121	0.70	2x2 factorial
10	Faulkner et al. (2015)	0.68	0.59	0.50	Two arm experiment
11	Broockman and Kalla (2016)	1	1.14	0.40	Two arm experiment
12	Gerber et al. (2008)*	0.465	0.485	0.00	Five arm experiment
13	Gerber et al. (2003)*	0.486	0.492	0.00	2x4 factorial

Table 1: Estimated reduction of variance of selected experiments in political science. Variance reduction is calculated with respect to the two most disparate experimental conditions in terms of variance and assumes uncorrelated potential outcomes. Asterisks indicate studies with binary outcome variables, though note that no artificial observations were added as in the simulations. Full citations in the Supplemental Materials.

improvements do not appear to be significantly related to the support of the outcome variable, be it binary or otherwise, as indicated by the asterisks in Table 1. Thus, the BANA method has significant potential for improving precision and, as shown by the simulation evidence, very little loss of power when the uniform design is optimal and the initial batch is large enough. Recall that reduction in variance is approximately proportional to reductions in necessary sample size (see Branson, Li and Ding, 2022, for some sample size calculations for finite-population causal inference).

7 Practical advice

We now discuss advice for how and when practitioners can make the best use of adaptive designs, and in particular BANA. First, in order for the BANA design to successfully reduce variance, the treatment and the units in both the first and second batch need to be similar. Specifically, the variance in each treatment arm needs to be the same or close between batches in order for the BANA design to optimally allocate units into treatment and control in the second batch. Therefore, BANA may

work best when run within a single experiment by splitting the sample into two batches, rather than when using a prior study on a different population or with a set of treatments that are not precisely the same as a first batch.

Second, a conservative approach to using BANA is to shrink the allocation probabilities towards uniform assignment, especially in settings where the first and second batch may differ. We developed one version of this approach for binary outcomes, but one could generalize this ideal to other types of outcomes. By weighting towards uniform allocation, one can partially offset potential harm from large changes in variance from batch to batch, for example if the first batch is temporally or spatially distant to the second batch. The disadvantage of this approach is that it will reduce the potential variance gains from using the BANA design.

Third, it is possible to extend the BANA approach to do multiple batches or to take an "online" approach. Here, the researcher would update the weights sequentially through the multiple batches as more data is collected. This may also help reduce dependence of the weights on an initial "poor" batch that does not resemble later units. Dai, Gradu and Harshaw (2023) discuss one such adaptive design and also implement adaptive constraints on how far the allocation probabilities can deviate from uniform.

Fourth, researchers can implement BANA in a block randomized design by simply applying the adaptive design within each block independently. Ravichandran et al. (2023) provide formal theory on optimal allocation in block randomized designs. Batch adaptive Neyman allocation could be particularly advantageous with block randomized designs if heterogeneity varies from block to block. For example, if there is a block of units who are all very homogeneous in terms of outcomes and another block that is very heterogenous in terms of treatment effects, the BANA design could take advantage of this to assign more uniform weights to the former compared to the latter.

Fifth, BANA will likely work best with continuous outcomes and should be used with some caution with binary outcomes. As demonstrated in our theoretical results and simulations, BANA can, in theory, provide improvement over uniform in binary settings, but the improvement is limited due to

constraints on the difference in variances in binary settings and the improvement may be erased by estimation variability. Therefore, we recommend using BANA in binary settings only when a large first batch can be used and the treatment effect is believed to be of reasonable size, which implies larger difference in variances in the binary setting. Luckily, continuous settings do not have the same constraints on the size of variance or variance being tied to means and treatment effect sizes. For binary settings in our simulations, we found a first batch size of at least 100–200 worked well whereas for the continuous outcome simulation even a small batch size of 25 leads to BANA being close to the true optimal allocation.

8 Conclusion

In this paper, we have shown how batch-adaptive experimental designs that leverage Neyman allocation can lead to improved statistical efficiency. Our proposed design is purposefully simple so that it can be implemented easily by practitioners, but there are more complex optimal designs that help to address different criteria. A number of these complexities should be the topic of future research in this area. In particular, incorporating the costs of recruitment into this analysis might provide an optimality that balances statistical criteria with the ability to implement the design. Furthermore, it would be beneficial for survey platforms to build this type of design into their products, allowing researchers to conduct these studies without manually adjusting the treatment allocations.

Beyond the efficiency gains of the BANA design, there are additional advantages to fielding experiments in batches. Pilot studies are already a common practice in experimental work, allowing researchers to find potential problems in their study design before the final study is conducted. The BANA design formalizes the use of the pilot study to inform the optimal design of the main study, but the initial batch can also be used to make the usual changes to the design—clarify prompts, change the measurement of variables, include additional moderating or mediating variables, and so on.

References

Armstrong, Timothy B. 2022. "Asymptotic Efficiency Bounds for a Class of Experimental Designs." *arXiv:2205.02726 [stat]* .

Berry, Donald A. and Bert Fristedt. 1985. *Bandit problems*. Dordrecht: Springer Netherlands.

Branson, Zach, Xinran Li and Peng Ding. 2022. "Power and Sample Size Calculations for Rerandomized Experiments." *arXiv preprint arXiv:2201.02486* .

Cochran, William G. 1977. *Sampling Techniques, 3rd Edition.* New York: John Wiley.

Cytrynbaum, Max. 2021. "Designing Representative and Balanced Experiments by Local Randomization." *arXiv:2111.08157 [econ, math, stat]* .

Dai, Jessica, Paula Gradu and Christopher Harshaw. 2023. "Clip-OGD: An Experimental Design for Adaptive Neyman Allocation in Sequential Experiments." *arXiv preprint arXiv:2305.17187* .

Dimmery, Drew. 2019. "Adaptive Neyman Allocation." Working paper.

Fisher, R.A. 1935. *The design of experiments.* Edinburgh: Oliver and Boyd.

Friedrich, Ulf, Ralf Münnich, Sven de Vries and Matthias Wagner. 2015. "Fast integer-valued algorithms for optimal allocations under constraints in stratified sampling." *Computational Statistics & Data Analysis* 92:1–12.

Hadad, Vitor, David A. Hirshberg, Ruohan Zhan, Stefan Wager and Susan Athey. 2021. "Confidence intervals for policy evaluation in adaptive experiments." *Proceedings of the National Academy of Sciences* 118(15):e2014602118.

Hahn, Jinyong, Keisuke Hirano and Dean Karlan. 2011. "Adaptive Experimental Design Using the Propensity Score." *Journal of Business & Economic Statistics* 29(1):96–108.

Howard, Steven R., Aaditya Ramdas, Jon McAuliffe and Jasjeet Sekhon. 2021. "Time-uniform, nonparametric, nonasymptotic confidence sequences." *The Annals of Statistics* 49(2).

Hu, Feifang and Li-Xin Zhang. 2004. "Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials." *The Annals of Statistics* 32(1).

Imai, Kosuke, Gary King and Elizabeth A. Stuart. 2008. "Misunderstandings between experimentalists and observationalists about causal inference." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(2):481–502.

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Behavioral Sciences*. Cambridge University Press.

Melfi, Vincent F., Connie Page and Margarida Geraldes. 2001. "An adaptive randomized design with application to estimation." *Canadian Journal of Statistics* 29(1):107–116.

Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97(4):558–625.

Offer-Westort, Molly, Alexander Coppock and Donald P. Green. 2021. "Adaptive Experimental Design: Prospects and Applications in Political Science." *American Journal of Political Science* 65(4):826–844.

Pashley, Nicole E and Luke W Miratrix. 2022. "Block what you can, except when you shouldn't." *Journal of Educational and Behavioral Statistics* 47(1):69–100.

Ravichandran, Arun, Nicole E Pashley, Brian Libgober and Tirthankar Dasgupta. 2023. "Optimal allocation of sample size for randomization-based inference from 2^K factorial designs." *arXiv preprint arXiv:2306.12394* .

Robbins, Herbert. 1952. "Some aspects of the sequential design of experiments." *Bulletin of the American Mathematical Society* 58(5):527–535.

Rosenman, Evan T. R. and Art B. Owen. 2021. "Designing experiments informed by observational studies." *Journal of Causal Inference* 9(1):147–171.

Rubin, Donald B. 1980. "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment." *J. Amer. Statist. Assoc.* 75(371):591–593.

Solomon, H and S Zacks. 1970. "Optimal Design of Sampling from Finite Populations: A Critical Review and Indication of New Research Areas." *Journal of the American Statistical Association* 65(330):653–677.

Sukhatme, P. V. 1935. "Contribution to the Theory of the Representative Method." *Supplement to the Journal of the Royal Statistical Society* 2(2):253.

Tabord-Meehan, Max. 2021. "Stratification Trees for Adaptive Randomization in Randomized Controlled Trials." *arXiv:1806.05127 [econ, stat]* .

Zhang, Kelly, Lucas Janson and Susan Murphy. 2020. Inference for Batched Bandits. In *Advances in Neural Information Processing Systems*, ed. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin. Vol. 33 Curran Associates, Inc. pp. 9818–9829.

Supplemental Materials (to appear online)

A Derivation of the Control-augmented Neyman Allocation

To show that control-augmented Neyman allocation is optimal, we set up the constrained optimization problem

$$\underset{(\pi_0,\pi_1,\ldots,\pi_J)}{\arg\min} \sum_{j=1}^{J} \left(\frac{S^2(j)}{\pi_j N} + \frac{S^2(0)}{\pi_0 N} \right) \quad \text{s.t.} \sum_{j=0}^{J} \pi_j = 1.$$

The first order conditions of this optimization problem imply that for $j > 0$

$$\frac{\pi_j}{\pi_0} = \frac{S(j)}{\sqrt{J} S(0)}.$$

The sum-to-one constraint implies further that

$$\begin{aligned} \pi_0 \left(\sum_{j=1}^{J} \frac{\pi_j}{\pi_0} \right) &= 1 - \pi_0, \\ \implies \pi_0 \frac{\sum_{j=1}^{J} S(j)}{\sqrt{J} S(0)} &= 1 - \pi_0 \end{aligned}$$

which after rearranging yields,

$$\pi_0 = \frac{\sqrt{J} S(0)}{\sqrt{J} S(0) + \sum_{j=1}^{J} S(j)}.$$

Substituting this into the first order conditions yields the remainder of the Neyman allocation.

B Proofs

B.1 Relative efficiency of the Neyman allocation

Define the finite sample correlation between the potential outcomes as

$$\rho = \frac{1}{(N-1)S(1)S(0)} \sum_{i=1}^{N} \left\{ Y_i(1) - \overline{Y}(1) \right\} \left\{ Y_i(0) - \overline{Y}(0) \right\},$$

and note that we have $S^2(1,0) = S^2(1) + S^2(0) - 2\rho S(1)S(0) = S^2(0)(1 + \delta^2 - 2\delta\rho)$. We can write the relative efficiency as

$$\frac{\mathbb{V}_{na}[\widehat{\tau}_1]}{\mathbb{V}_{u}[\widehat{\tau}_1]} = \frac{\{S(1) + S(0)\}^2 - S^2(1,0)}{2S^2(1) + 2S^2(0) - S^2(1,0)} = 1 - \frac{\{S(1) - S(0)\}^2}{2S^2(1) + 2S^2(0) - S^2(1,0)}.$$

Substituting in $S(1) = \delta S(0)$ yields,

$$\begin{aligned}\frac{\mathbb{V}_{na}[\widehat{\tau}_1]}{\mathbb{V}_u[\widehat{\tau}_1]} &= \frac{S^2(0)(1+\delta)^2 - S^2(0)(1+\delta^2-2\delta\rho)}{2S^2(0)(1+\delta^2) - S^2(0)(1+\delta^2-2\delta\rho)} \\ &= \frac{2\delta(1+\rho)}{1+\delta^2+2\delta\rho}.\end{aligned}$$

The calculation under the superpopulation model differs slightly. Let $\sigma^2(1) = \mathbb{V}^P[Y_i(1)]$ and $\sigma^2(0) = \mathbb{V}^P[Y_i(0)]$ be the (super)population variances of the potential outcomes, where we use $\mathbb{V}^P$ to denote the variance over both randomization and random sampling from the population. We define the relative standard deviations similarly to the finite sample case as $\delta = \sigma(1)/\sigma(0)$. Let $\mathbb{V}^P_{na}[\widehat{\tau}_1]$ and $\mathbb{V}^P_u[\widehat{\tau}_1]$ be the superpopulation variances of the estimators. Standard experimental design results give

$$\begin{aligned}\mathbb{V}^P_u[\widehat{\tau}_1] &= \frac{2(\sigma^2(1)+\sigma^2(0))}{N} = \frac{2(\delta^2\sigma^2(0)+\sigma^2(0))}{N}, \\ \mathbb{V}^P_{na}[\widehat{\tau}_1] &= \frac{(\sigma(1)+\sigma(0))^2}{N} = \frac{(\delta\sigma(0)+\sigma(0))^2}{N}.\end{aligned}$$

From this, we can see the superpopulation relative efficiency is

$$\frac{\mathbb{V}^P_{na}[\widehat{\tau}_1]}{\mathbb{V}^P_u[\widehat{\tau}_1]} = \frac{(1+\delta)^2}{2(\delta^2+1)}.$$

B.2 Unbiasedness

In this section we prove the unbiasedness of the stratified treatment effect estimator and the conservativeness of the stratified variance estimator (and unbiased under constant effects for the latter). For the purposes of showing unbiasedness of $\widehat{\tau}_s$, is sufficient to show that $\overline{Y}_2^{\text{obs}}(d)$ is unbiased for $\overline{Y}_2(d)$ because $\overline{Y}_1^{\text{obs}}(d)$ is unbiased for $\overline{Y}_1(d)$ for all d by standard results on experimental design. We focus here on the case with a completely randomized design and take a finite population point of view so that the only source of randomness comes from the treatment assignment. Let $\boldsymbol{D}_1 = (D_1, \ldots, D_{N_1})$

be the realized assignments in the first batch. Then using the law of iterated expectations we have:

$$\begin{aligned}
\mathbb{E}\left[\overline{Y}^{\text{obs}}(d)\right] &= \sum_{i=N_1+1}^{N_2} Y_i(d)\mathbb{E}\left[\frac{1}{n_{2d}}\mathbb{I}(D_i = d)\right] \\
&= \sum_{i=N_1+1}^{N_2} Y_i(d)\mathbb{E}\left[\frac{1}{n_{2d}}\mathbb{E}\left\{\mathbb{I}(D_i = d) \mid \boldsymbol{D}_1\right\}\right] \\
&= \sum_{i=N_1+1}^{N_2} Y_i(d)\mathbb{E}\left[\frac{1}{n_{2d}}\frac{n_{2d}}{N_2}\right] \\
&= \overline{Y}_2(d)
\end{aligned}$$

The second equality uses the fact that conditional on the draws in the first batch, the number of treated units in the second batch is fixed and the third equality is based on the design of the second batch.

Thus, we have

$$\begin{aligned}
\mathbb{E}[\widehat{\tau}_s(d,d')] &= \left(\frac{N_1}{N}\right)\mathbb{E}[\widehat{\tau}_1(d,d')] + \left(\frac{N_2}{N}\right)\mathbb{E}[\widehat{\tau}_2(d,d')] \\
&= \left(\frac{N_1}{N}\right)\left(\overline{Y}_1(d) - \overline{Y}_1(d')\right) + \left(\frac{N_2}{N}\right)\left(\overline{Y}_2(d) - \overline{Y}_2(d')\right) \\
&= \overline{Y}(d) - \overline{Y}(d')
\end{aligned}$$

Note that this unbiasedness holds regardless of whether or not the treatment effect varies between the two batches and the proof does not depend on the specific allocation in the second batch.

B.3 Variance

Let $\sigma_1^2(d,d') = \mathbb{V}[\widehat{\tau}_1(d,d')]$ and $\sigma_2^2(d,d') = \mathbb{V}[\widehat{\tau}_2(d,d') \mid \boldsymbol{D}_1]$, where

$$\sigma_2^2(d,d') = \frac{S_2^2(d)}{n_{2d}} + \frac{S_2^2(d')}{n_{2d'}} - \frac{S_2^2(d,d')}{N_2}.$$

Using the law of total variance, we have

$$\begin{aligned}
\mathbb{V}[\widehat{\tau}_s(d,d')] &= \mathbb{E}\left[\mathbb{V}\left(\widehat{\tau}_s(d,d')|\boldsymbol{D}_1\right)\right] + \mathbb{V}\left(\mathbb{E}\left[\widehat{\tau}_s(d,d')|\boldsymbol{D}_1\right]\right) \\
&= \mathbb{E}\left[\mathbb{V}\left(\frac{N_2}{N}\left(\widehat{\tau}_2(d,d')\right)|\boldsymbol{D}_1\right)\right] + \mathbb{V}\left(\frac{N_1}{N}\widehat{\tau}_1(d,d') + \mathbb{E}\left[\frac{N_2}{N}\widehat{\tau}_2(d,d')|\boldsymbol{D}_1\right]\right) \\
&= \mathbb{E}\left[\frac{N_2^2}{N^2}\sigma_2^2(d,d')\right] + \mathbb{V}\left(\frac{N_1}{N}\widehat{\tau}_1(d,d')\right) \\
&= \frac{N_2^2}{N^2}\mathbb{E}\left[\sigma_2^2(d,d')\right] + \frac{N_1^2}{N^2}\sigma_1^2(d,d')
\end{aligned}$$

Now, we can turn to the variance estimator. By the usual results on variance estimators in completely randomized designs and the law of iterated expectations, we have:

$$\begin{aligned}\mathbb{E}\left[\widehat{V}[\widehat{\tau}_s(d,d')]\right] &= \left(\frac{N_1}{N}\right)^2\left(\frac{S_1^2(d)}{N_{1d}}+\frac{S_1^2(d')}{N_{1d'}}\right)+\left(\frac{N_2}{N}\right)^2\mathbb{E}\left[\mathbb{E}[\widehat{\mathbb{V}}[\widehat{\tau}_2(d,d')] \mid \boldsymbol{D}_1]\right] \\ &= \left(\frac{N_1}{N}\right)^2\left(\frac{S_1^2(d)}{N_{1d}}+\frac{S_1^2(d')}{N_{1d'}}\right)+\left(\frac{N_2}{N}\right)^2\mathbb{E}\left[\frac{S_2^2(d)}{n_{2d}}+\frac{S_2^2(d')}{n_{2d'}}\right].\end{aligned}$$

Thus, the bias of our estimator is

$$\mathbb{E}\left[\widehat{V}[\widehat{\tau}_s(d,d')]\right]-\mathbb{V}[\widehat{\tau}_s(d,d')] = \left(\frac{N_1}{N}\right)^2\frac{S_1^2(d,d')}{N_{1d}+N_{1d'}}+\left(\frac{N_2}{N}\right)^2\mathbb{E}\left[\frac{S_2^2(d,d')}{n_{2d}+n_{2d'}}\right]\geq 0$$

Thus, we can see that our variance estimator will be conservative in the sense that it has a positive bias unless there is no variation in the treatment effect across units in which case the variance estimator is unbiased.

C Additional Simulation Results

C.1 Gaussian outcome

Here we present additional simulation results with Gaussian outcomes using either the full sample or only the second batch to estimate effects, respectively.

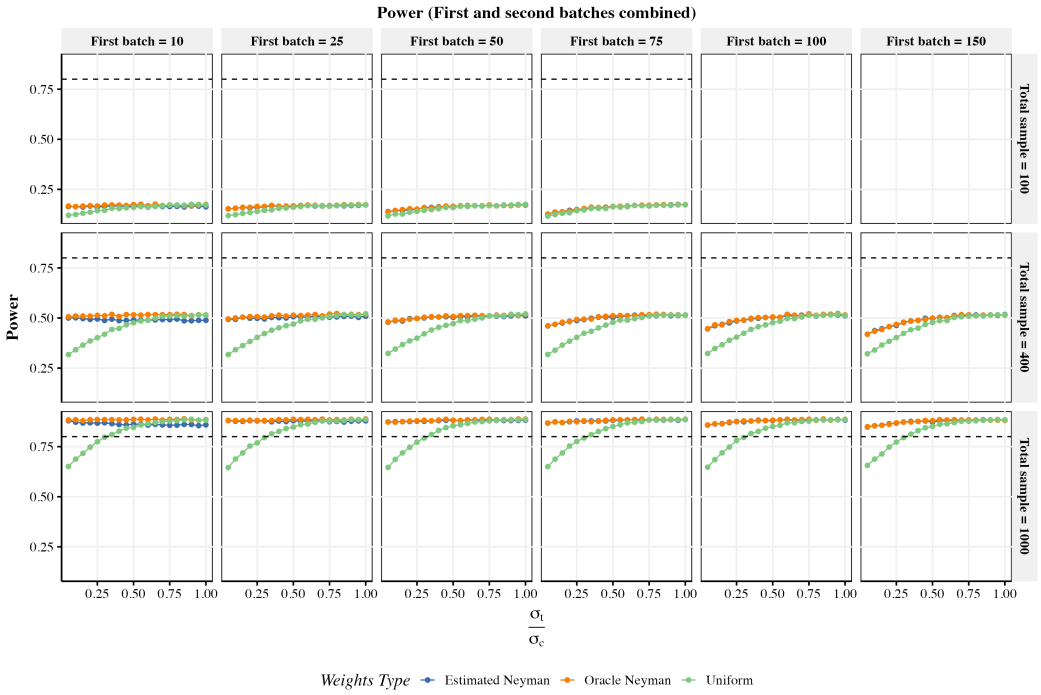

Figure SM.3: Simulation results utilizing both batches, Gaussian outcome. Effects are estimated using the stratified estimator. Panels differ horizontally by the size of the first batch and vertically by the total sample size. The x-axis displays the ratio of the standard deviations of the outcome in the treatment and control conditions, but note that overall variability of the outcome is constant. Colors indicate allocation scheme. The outcome is power to detect the treatment effect. Instances where the initial batch size is greater than or equal to total sample size are omitted. Estimated Neyman weights converge to Oracle Neyman weights at initial batch sizes around 25–50 suggesting no need to allocate more than 50 units to the initial batch.

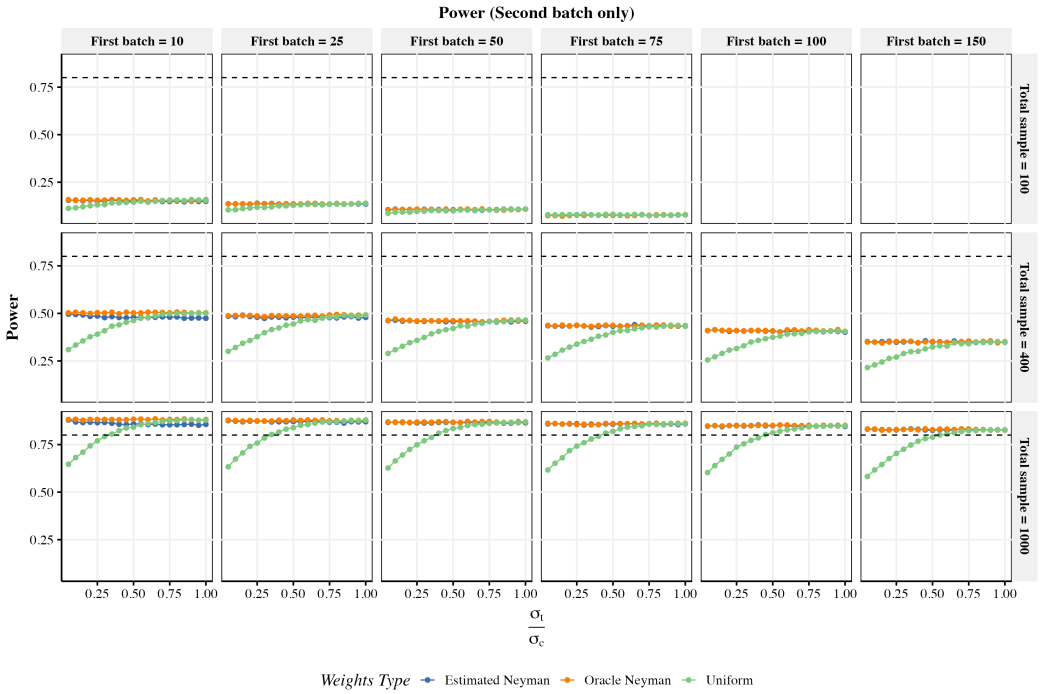

Figure SM.4: Simulation results utilizing the second batch only, Gaussian outcome. Effects are estimated using the traditional difference in means estimator since there is no longer dependence between units. Panels differ horizontally by the size of the first batch and vertically by the total sample size. The x-axis displays the ratio of the standard deviations of the outcome in the treatment and control conditions, but note that overall variability of the outcome is constant. Colors indicate allocation scheme. The outcome is power to detect the treatment effect. Instances where the initial batch size is greater than or equal to total sample size are omitted. Estimated Neyman weights converge to Oracle Neyman weights at initial batch sizes around 25–50 suggesting no need to allocate more than 50 units to the initial batch.

C.2 Binary outcome

Here we present additional simulation results with binary outcomes. First, we display the impact of additional observations at the grand mean before estimating Neyman weights as described in the main text. Then, we investigate how power can be improved by the BANA design in the binary case. Figures SM.7 and SM.8 display the power to detect the specified treatment effect of 0.1 under each of the designs using either both the first and second batch or the second batch only respectively.

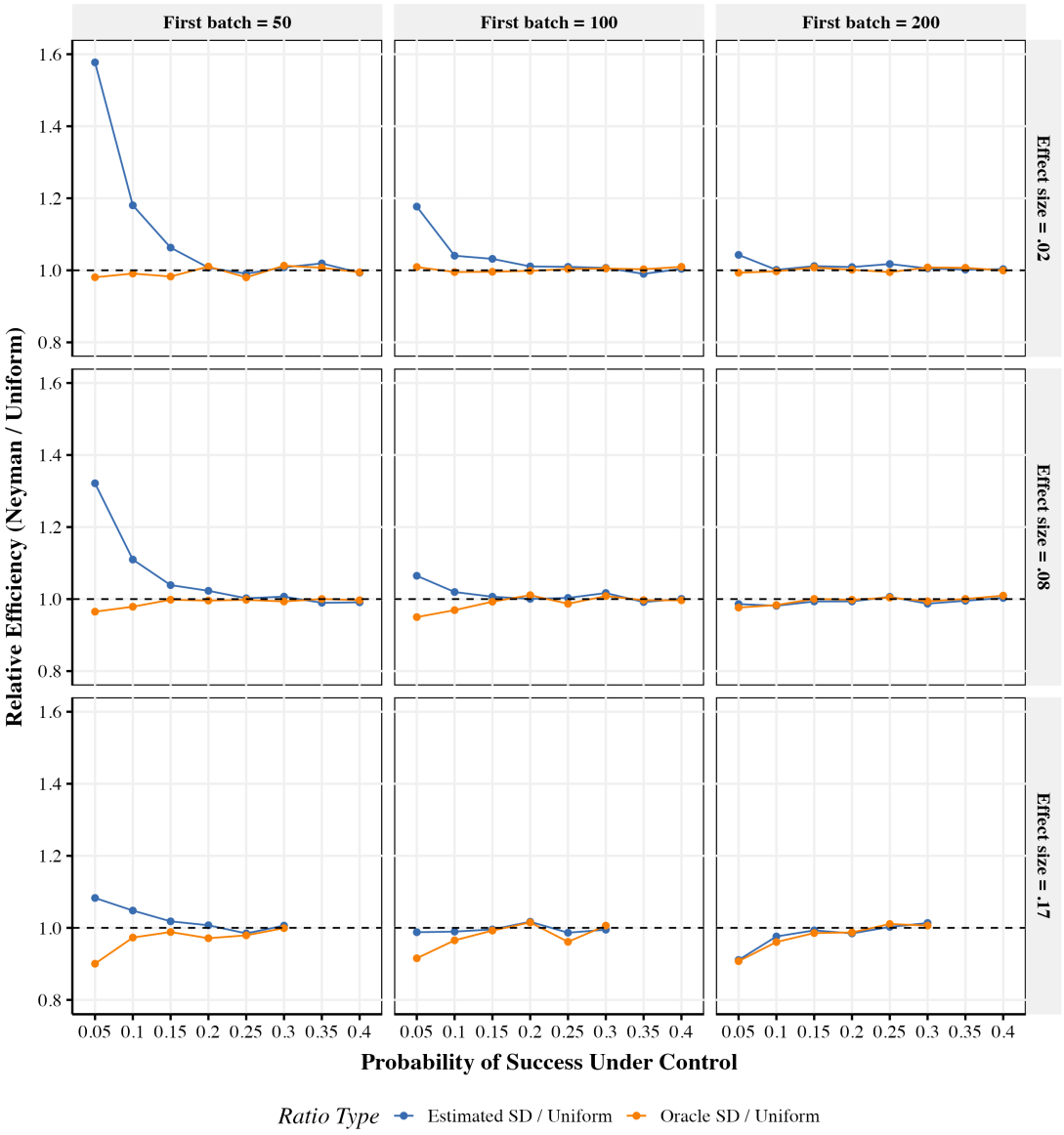

Figure SM.5: Binary outcome simulation results. $N = 1,000$, $R = 30,000$. Horizontal panels differ by the size of the first batch and vertical panels differ by the effect size, informed by the meta-analysis below. The x-axis displays the probability of success in the control condition; note that unlike the Gaussian case, total variability is not constant. Colors indicate allocation scheme relative to uniform. Minimum allocation to each condition in the second batch is set to 0.1 and no additional observations were added to each condition in the first batch.

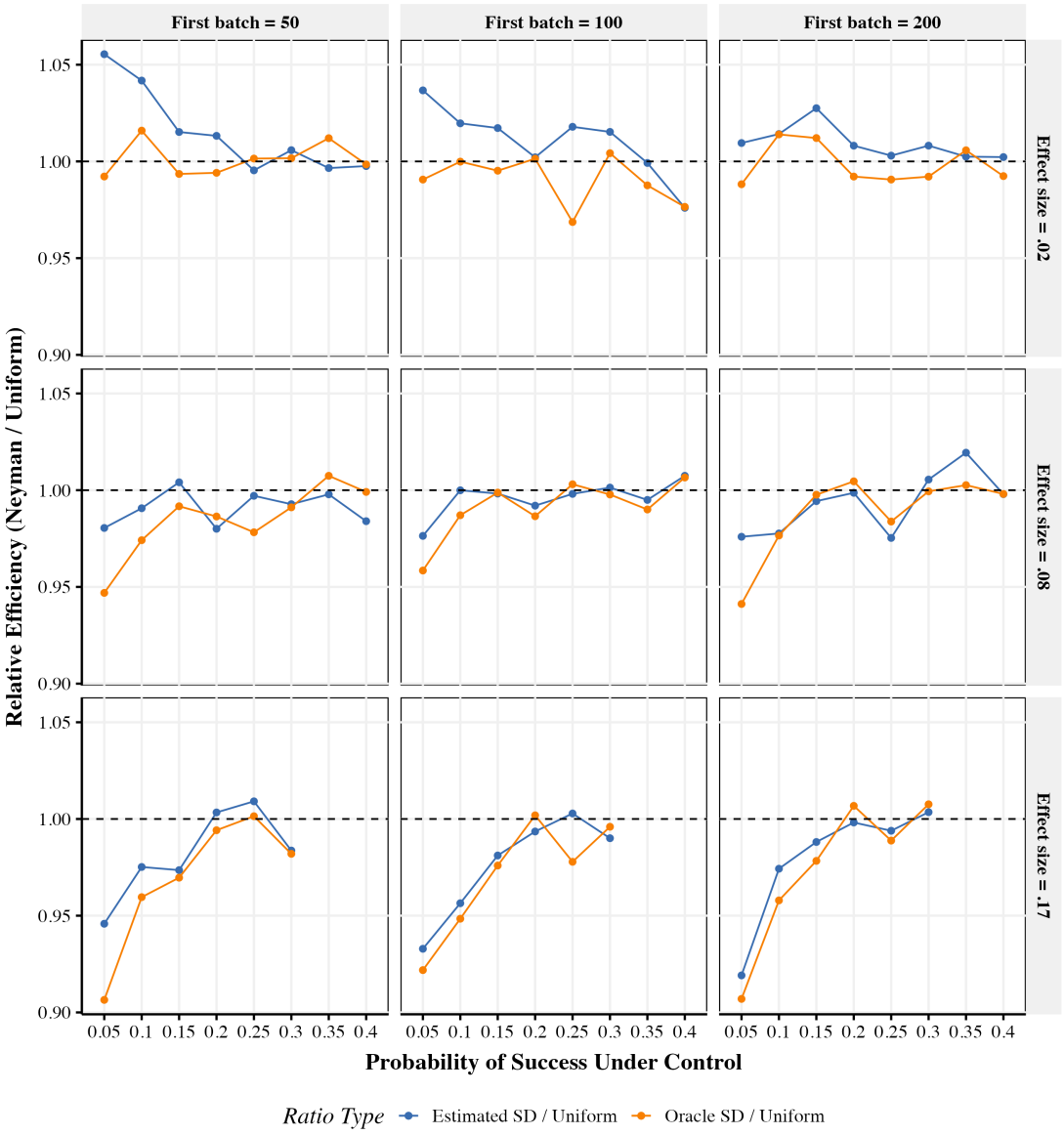

Figure SM.6: Binary outcome simulation results. $N = 1,000$, $R = 30,000$. Horizontal panels differ by the size of the first batch and vertical panels differ by the effect size, informed by the meta-analysis below. The x-axis displays the probability of success in the control condition; note that unlike the Gaussian case, total variability is not constant. Colors indicate allocation scheme relative to uniform. Minimum allocation to each condition in the second batch is set to 0.1 and ten additional observations were added to each condition in the first batch.

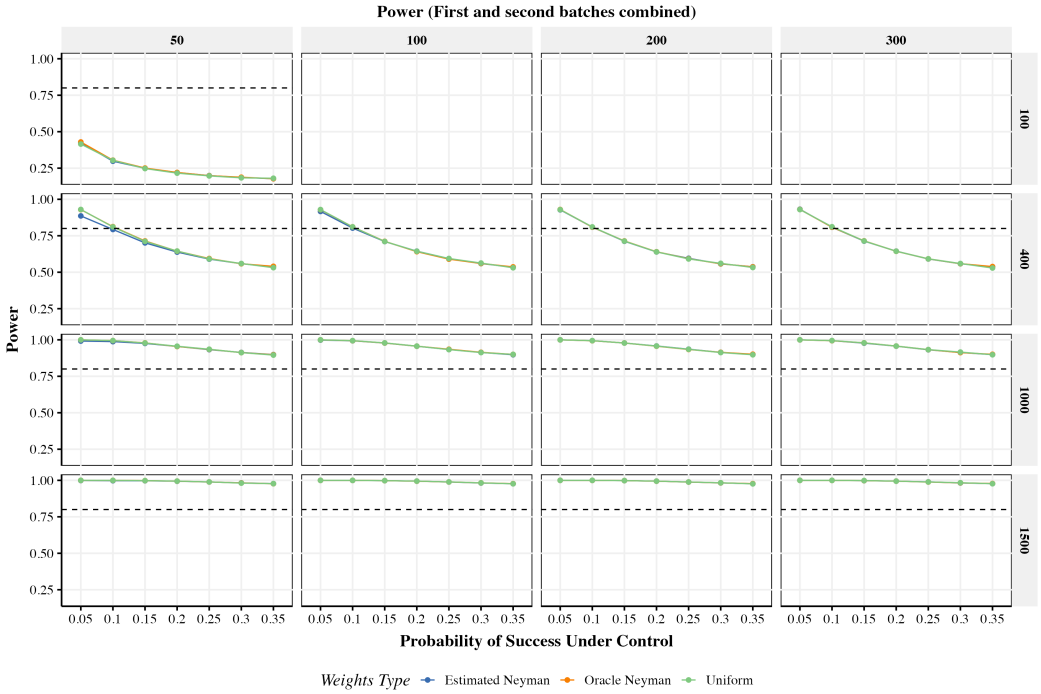

Figure SM.7: Simulation results utilizing both batches, binary outcome. $R = 10,000$. Effects are estimated using the stratified estimator. Panels differ horizontally by the size of the first batch and vertically by the total sample size. The x-axis displays the probability of success in the control condition, with the probability of success in treatment determined by a constant treatment effect of 0.1. Colors indicate allocation scheme. The outcome is power to detect the treatment effect. Instances where the initial batch size is greater than or equal to total sample size are omitted. Minimum allocation to each condition in the second batch is set to 0.1 and no additional observations were added to each condition in the first batch.

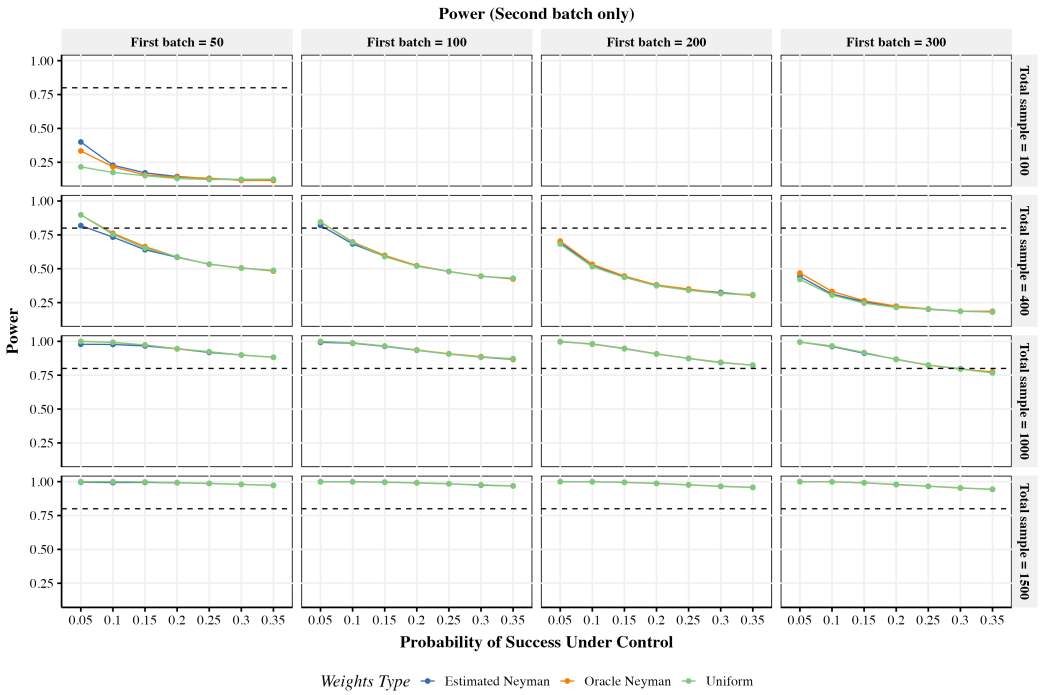

Figure SM.8: Simulation results utilizing the second batch only, binary outcome. $R = 10,000$. Effects are estimated using the traditional difference in means estimator since there is no longer dependence between units. Panels differ horizontally by the size of the first batch and vertically by the total sample size. The x-axis displays the probability of success in the control condition, with the probability of success in treatment determined by a constant treatment effect of 0.1. Colors indicate allocation scheme. The outcome is power to detect the treatment effect. Instances where the initial batch size is greater than or equal to total sample size are omitted. Estimated Neyman weights converge to Oracle Neyman weights at initial batch sizes around 25-50 suggesting no need to allocate more than 50 units to the initial batch. Minimum allocation to each condition in the second batch is set to 0.1 and no additional observations were added to each condition in the first batch.

Articles in Meta-analysis

Broockman, David. 2013. "Black politicians are more intrinsically motivated to advance blacks' interests: A field experiment manipulating political incentives." *American Journal of Political Science* 57(3):521–536.

Broockman, David and Joshua Kalla. 2016. "Durably reducing transphobia: A field experiment on door-to-door canvassing." *Science* 352(6282):220–224.

De Vreese, Claes. 2004. "The effects of strategic news on political cynicism, issue evaluations, and policy support: A two-wave experiment." *Mass Communication & Society* 7(2):191–214.

Eble, Alex, Chris Frost, Alpha Camara, Baboucarr Bouy, Momodou Bah, Maitri Sivaraman, Pei-Tseng Jenny Hsieh, Chitra Jayanty, Tony Brady, Piotr Gawron et al. 2021. "How much can we remedy very low learning levels in rural parts of low-income countries? Impact and generalizability of a multi-pronged para-teacher intervention from a cluster-randomized trial in The Gambia." *Journal of Development Economics* 148:102539.

Eggers, Andrew C, Nick Vivyan and Markus Wagner. 2018. "Corruption, accountability, and gender: Do female politicians face higher standards in public life?" *The Journal of Politics* 80(1):321–326.

Faulkner, Nicholas, Aaron Martin and Kyle Peyton. 2015. "Priming political trust: Evidence from an experiment." *Australian Journal of Political Science* 50(1):164–173.

Gerber, Alan S., Donald P. Green and Christopher W. Larimer. 2008. "Social pressure and voter turnout: Evidence from a large-scale field experiment." *American political Science review* 102(1):33–48.

Gerber, Alan S., Donald P. Green and Ron Shachar. 2003. "Voting may be habit-forming: evidence from a randomized field experiment." *American journal of political science* 47(3):540–550.

Goff, Sandra H, Timothy M Waring and Caroline L Noblet. 2017. "Does pricing nature reduce monetary support for conservation?: evidence from donation behavior in an online experiment." *Ecological economics* 141:119–126.

Holman, Mirya R, Jennifer L Merolla and Elizabeth J Zechmeister. 2016. "Terrorist threat, male stereotypes, and candidate evaluations." *Political Research Quarterly* 69(1):134–147.

Siegel, Alexandra A and Vivienne Badaan. 2020. "# No2Sectarianism: Experimental approaches to reducing sectarian hate speech online." *American Political Science Review* 114(3):837–855.

Simas, Elizabeth N and Doug Murdoch. 2020. ""I Didn't Lie, I Misspoke": Voters' Responses to Questionable Campaign Claims." *Journal of Experimental Political Science* 7(2):75–88.

Vernby, Kåre and Rafaela Dancygier. 2019. "Can immigrants counteract employer discrimination? A factorial field experiment reveals the immutability of ethnic hierarchies." *PloS one* 14(7):e0218044.