An evaluation framework for personalization strategy experiment designs

C. H. Bryan Liu
bryan.liu12@imperial.ac.uk
Imperial College London & ASOS.com, UK

Emma J. McCoy
Imperial College London, UK

ABSTRACT

Online Controlled Experiments (OCEs) are the gold standard in evaluating the effectiveness of changes to websites. An important type of OCE evaluates different personalization strategies, which present challenges in low test power and lack of full control in group assignment. We argue that getting the right experiment setup — the allocation of users to treatment/analysis groups — should take precedence of post-hoc variance reduction techniques in order to enable the scaling of the number of experiments. We present an evaluation framework that, along with a few rule of thumbs, allow experimenters to quickly compare which experiment setup will lead to the highest probability of detecting a treatment effect under their particular circumstance.

KEYWORDS

Experimentation, Experiment design, Personalization strategies, A/B testing.

1 INTRODUCTION

The use of Online Controlled Experiments (OCEs, e.g. A/B tests) has become popular in measuring the impact of products and guiding business decisions on the Web. Major companies report running thousands of OCEs on any given day and many startups exist purely to manage OCEs. A large number of OCEs address simple variations on elements of the user experience based on random splits, e.g. showing a different colored button to users based on a user ID hash bucket. Here, we are interested in experiments that compare *personalization strategies*, complex sets of targeted customer interactions that are common in e-commerce and digital marketing. Examples of personalization strategies include the scheduling, budgeting and ordering of marketing activities directed at a user based on their purchase history.

Experiments for personalization strategies face two unique challenges. Firstly, strategies are often only applicable to a small fraction of the user base, and thus many simple experiment designs suffer from either a lack of sample size / statistical power, or diluted metric movement by including irrelevant samples [2]. Secondly, as users are not randomly assigned *a priori*, but must qualify to be treated with a strategy via their actions or attributes, groups of users subjected to different strategies cannot be assumed to be statistically equivalent and hence are not directly comparable.

While there are a number of variance reduction techniques (including stratification and control variates [3, 7]) that partially address the challenges, the strata and control variates involved can vary dramatically from one personalization strategy experiment to another, requiring many *ad hoc* adjustments. As a result, such techniques may not scale well when organizations design and run hundreds or thousands of experiments at any given time.

We argue that personalization strategy experiments should focus on the assignment of users from the strategies they qualified for to the treatment/analysis groups. We call this mapping process an *experiment setup*. Identifying the best experiment setup increases the chance to detect any treatment effect. An experimentation framework can also reuse and switch between different setups quickly with little custom input, ensuring the operation can scale. More importantly, the process does not hinder the subsequent application of variance reduction techniques, meaning that we can still apply the techniques if required.

To date, many experiment setups exist to compare personalization strategies. An increasingly popular approach is to compare the strategies using multiple control groups — Quantcast calls it a dual control [1], and Facebook calls it a multi-cell lift study [5]. In the two-strategy case, this involves running two experiments on two random partitions of the user base in parallel, with each experiment further splitting the respective partition into treatment/control and measuring the incrementality (the change in a metric as compared to the case where nothing is done) of each strategy. The incrementality of the strategies are then compared against each other.

Despite the setup above gaining traction in display advertising, there is a lack of literature on whether it (or any other candidate) is a good setup — one that has a higher sensitivity and/or apparent effect size than other setups. While Liu et al. [5] noted that multi-cell lift studies require a large number of users, they did not discuss how the number compares to other setups.[1] The ability to identify and adopt a better experiment setup can reduce the required sample size, and hence enable more cost-effective experimentation.

We address the gap in the literature by introducing an evaluation framework that compares experiment setups given two personalization strategies. The framework is designed to be flexible so that it is able to deal with a wide range of baselines and changes in user responses presented by any pairs of strategies (*situations* hereafter). We also recognize the need to quickly compare common setups, and provide some rule of thumbs on situations where a setup will be better than another. In particular, we outline the conditions where employing a multi-cell setup, as well as metric dilution, is desirable.

To summarize, our contributions are: (i) We develop a flexible evaluation framework for personalization strategy experiments, where one can compare two experiment setups given the situation presented by two competing strategies (Sec. 2); (ii) We provide simple rule of thumbs to enable experimenters who do not require the full flexibility of the framework to quickly compare common setups (Sec. 3); and (iii) We make our results useful to practitioners by making the code used in the paper (Sec. 4) publicly available.[2]

[1] A single-cell lift study is often used to measure the incrementality of a single personalization strategy, and hence is not a representative comparison.

[2] Code/supp. documents available on: github.com/liuchbryan/experiment_design_evaluation.

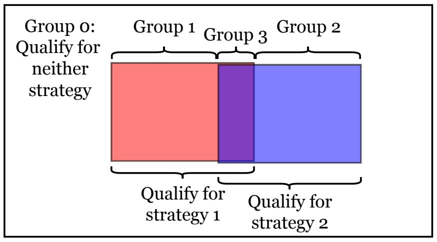

Figure 1: Venn diagram of the user groups in our evaluation framework. The outer, left inner (red), and right inner (blue) boxes represent the entire user base, those who qualify for strategy 1, and those who qualify for strategy 2 respectively.

2 EVALUATION FRAMEWORK

We first present our evaluation framework for personalization strategy experiments. The experiments compare two personalization strategies, which we refer to as strategy 1 and strategy 2. Often one of them is the existing strategy, and the other is a new strategy we intend to test and learn from. In this section we introduce (i) how users qualifying themselves into strategies creates non-statistically equivalent groups, (ii) how experimenters usually assign the users, and (iii) when we would consider an assignment to be better.

2.1 User grouping

As users qualify themselves into the two strategies, four disjoint groups emerge: those who qualify for neither strategy, those who qualify only for strategy 1, those who qualify only for strategy 2, and those who qualify for both strategies. We denote these groups (user) groups 0, 1, 2, and 3 respectively (see Fig. 1). It is perhaps obvious that we cannot assume those in different user groups are statistically equivalent and compare them directly.

We assume groups 0, 1, 2, 3 have n_0, n_1, n_2, and n_3 users respectively. We also assume the metric has a different distribution between groups, and within the same group, between the scenario where the group is subjected to the treatment associated to the corresponding strategy and where nothing is done (baseline). We list all group-scenario combinations in Table 1, and denote the mean and variance of the metric (μ_G, σ_G^2) for a combination G.[3]

2.2 Experiment setups

Many experiment setups exist and are in use in different organizations. Here we introduce four common setups of various sophistication, which we also illustrate in Fig. 2.

Setup 1 (Users in the intersection only). The setup considers users who qualify for both strategies only. The said users are randomly split (usually 50/50) into two (analysis) groups A and B, and are prescribed the treatment specified by strategies 1 and 2 respectively. The setup is easy to implement, though it is difficult to translate any learnings obtained from the setup to other user groups (e.g. those who qualify for one strategy only) [4].

Setup 2 (All samples). The setup is a simple A/B test where it considers all users, regardless on whether they qualify for any strategy or not. The users are randomly split into two analysis groups A and B, and are prescribed the treatment specified by strategy 1(2) if (i) they qualify under the strategy and (ii) they are in analysis group A(B). This setup is easiest to implement but usually suffers severely from a dilution in metric [2].

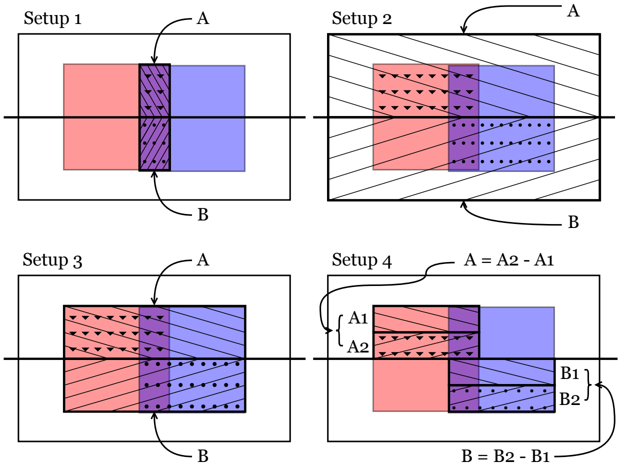

Figure 2: Experiment setups overlaid on the user grouping Venn diagram in Fig. 1. The hatched boxes indicate who are included in the analysis, and the downward triangles and dots indicate who are subjected to treatment prescribed under strategies 1 and 2 respectively. See Sec. 2.2 for a detailed description.

Setup 3 (Qualified users only). The setup is similar to Setup 2 except only those who qualified for at least one strategy ("triggered" users in some literature [2]) are included in the analysis groups. The setup sits between Setup 1 and Setup 2 in terms of user coverage, and has the advantage of capturing the most number of useful samples yet having the least metric dilution. However, the setup also prevents one from telling the incrementality of a strategy itself, but only the difference in incrementalities between two strategies.

Setup 4 (Dual control / multi-cell lift test). As described in Section 1, the setup first split the users randomly into two randomization groups. For the first randomization group, we consider those who qualify for strategy 1, and split them into analysis groups $A1$ and $A2$. Group $A2$ receives the treatment prescribed under strategy 1, and group $A1$ acts as control. The incrementality for strategy 1 is then the difference in metric between groups $A2$ and $A1$. We apply the same process to the second randomization group, with strategy 2 and analysis groups $B1$ and $B2$ in place, and compare the incrementality for strategies 1 and 2. The setup allows one to obtain the incrementality of each individual strategy and minimizes metric dilution. Though it also leaves a number of samples unused and creates extra analysis groups, and hence generally suffers from a low test power [5].

2.3 Evaluation criteria

There are a number of considerations when one evaluates competing experiment setups. They range from technical considerations (e.g. the complexity of setting up the setups) to business considerations (e.g. if the incrementality of individual strategies is required).

[3] For example, the metric for Group 1 without any interventions has mean and variance $(\mu_{C1}, \sigma_{C1}^2)$, and that for Group 2 with the treatment prescribed under strategy 2 has mean and variance $(\mu_{I2}, \sigma_{I2}^2)$.

	Group 0	Group 1	Group 2	Group 3
Baseline (**C**ontrol)	$C0$	$C1$	$C2$	$C3$
Under treatment (**I**ntervention)	/	$I1$	$I2$	Under strategy 1: Iϕ Under strategy 2: Iψ

Table 1: All group-scenario combinations in our evaluation framework for personalization strategy experiments. The columns represent the groups described in Fig. 1. The baseline represents the scenario where nothing is done. We assume those who qualify for both strategies (Group 3) can only receive treatment(s) associated to either strategy.

Here we focus on the statistical aspect and propose two evaluation criteria: (i) the actual effect size of a treatment as presented by the two analysis groups in an experiment setup, and (ii) the sensitivity of the experiment represented by the minimum detectable effect (MDE) under a pre-specified test power. Both criteria are necessary as the former indicates whether a setup suffers from metric dilution, whereas the latter indicates whether the setup suffers from lack of power/sample size. An ideal setup should yield a high actual effect size and a high sensitivity (i.e. a low MDE),[4] though as we observe in the next section it is usually a trade-off.

We formally define the two evaluation criteria from first principles while introducing relevant notations along the way. Let A and B be the two analysis groups in an experiment setup, with user responses randomly distributed with mean and variance (μ_A, σ_A^2) and (μ_B, σ_B^2) respectively. We first recall that if there are sufficient samples, the sample mean of the two groups approximately follows the normal distribution by the Central Limit Theorem:

$$\bar{A} \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu_A, \sigma_A^2/n_A\right), \quad \bar{B} \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu_B, \sigma_B^2/n_B\right), \tag{1}$$

where n_A and n_B are the number of samples taken from A and B respectively. The difference in the sample means then also approximately follows a normal distribution:

$$\bar{D} \triangleq (\bar{B} - \bar{A}) \overset{\text{approx.}}{\sim} \mathcal{N}\left(\Delta \triangleq \mu_B - \mu_A,\ \sigma_{\bar{D}}^2 \triangleq \sigma_A^2/n_A + \sigma_B^2/n_B\right). \tag{2}$$

Here, Δ is the actual effect size that we are interested in.

The definition of the MDE θ^* requires a primer to the power of a statistical test. A common null hypothesis statistical test in personalization strategy experiments uses the two-tailed hypotheses $H_0 : \Delta = 0$ and $H_1 : \Delta \neq 0$, with the test statistic under H_0 being:

$$T \triangleq \overline{D}/\sigma_{\bar{D}} \overset{\text{approx.}}{\sim} \mathcal{N}(0, 1). \tag{3}$$

We recall the null hypothesis will be rejected if $|T| > z_{1-\alpha/2}$, where α is the significance level and $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of a standard normal. Under a *specific* alternate hypothesis $\Delta = \theta$, the power is specified as

$$1 - \beta_\theta \triangleq \Pr\left(|T| > z_{1-\alpha/2} \,|\, \Delta = \theta\right) \approx 1 - \Phi\left(z_{1-\alpha/2} - |\theta|/\sigma_{\bar{D}}\right). \tag{4}$$

where Φ denotes the cumulative density function of a standard normal.[5] To achieve a minimum test power $\pi_{\min}$, we require that $1 - \beta_\theta > \pi_{\min}$. Substituting Eq. (4) into the inequality and rearranging to make θ the subject yields the effect sizes that the test will be able to detect with the specified power:

$$|\theta| > (z_{1-\alpha/2} - z_{1-\pi_{\min}})\, \sigma_{\bar{D}}. \tag{5}$$

θ^* is then defined as the positive minimum θ that satisfies Ineq. (5), i.e. that specified by the RHS of the inequality.

We finally define what it means to be better under these evaluation criteria. WLOG we assume the actual effect size of the two competing experiment setups are positive,[6] and say a setup S is superior to another setup R if, all else being equal,

(i) S produces a higher actual effect size ($\Delta_S > \Delta_R$) *and* a lower minimum detectable effect size ($\theta_S^* < \theta_R^*$), or
(ii) The gain in actual effect is greater than the loss in sensitivity:

$$\Delta_S - \Delta_R > \theta_S^* - \theta_R^*, \tag{6}$$

which means an actual effect still stands a higher chance to be observed under S.

3 COMPARING EXPERIMENT SETUPS

Having described the evaluation framework above, in this section we use the framework to compare the common experiment setups described in Sec. 2.2. We will first derive the actual effect size and MDE in Sec. 3.1, and using the result to create rule of thumbs on (i) whether diluting the metric by including users who qualify for neither strategies is beneficial (Sec. 3.2) and (ii) if dual control is a better setup for personalization strategy experiments (Sec. 3.3), two questions that are often discussed among e-commerce and marketing-focused experimenters. For brevity, we relegate most of the intermediate algebraic work when deriving the actual & minimum detectable effect sizes, as well as the conditions that lead to a setup being superior, to our supplementary document.[2]

3.1 Actual & minimum detectable effect sizes

We first present the actual effect size and MDE of the four experiment designs. For each setup we first compute the sample size, metric mean, and metric variance of the analysis groups (here we present only one of them),[7] which arises as a mixture of user groups described in Sec. 2.1. We then substitute the quantities computed into the definitions of Δ (see Eq. (2)) and θ^* (see Ineq. (5)) to obtain the setup-specific actual effect size and MDE. We assume all random splits are done 50/50 in these setups to maximize the test power.

3.1.1 Setup 1 (Users in the intersection only). The setup randomly splits user group 3 into two analysis groups, each with $n_3/2$ samples. Users in analysis group A are provided treatment under strategy 1, and hence the group metric has a mean and variance of $(\mu_{I\phi}, \sigma_{I\phi}^2)$. The actual effect size and MDE for Setup 1 are hence:

$$\Delta_{S1} = \mu_{I\psi} - \mu_{I\phi}, \tag{7}$$

$$\theta_{S1}^* = (z_{1-\alpha/2} - z_{1-\pi_{\min}})\sqrt{2(\sigma_{I\phi}^2 + \sigma_{I\psi}^2)/n_3}. \tag{8}$$

[4] We will use the terms "high(er) sensitivity" and "low(er) MDE" interchangeably.

[5] The approximation in Eq. (4) is tight for experiment design purposes, where $\alpha < 0.2$ and $1 - \beta > 0.6$ for nearly all cases.

[6] If both the actual effect sizes are negative, we simply swap the analysis groups. If the actual effect sizes are of opposite signs, it is likely an error.

[7] Expressions for other analysis groups can be easily obtained by substituting in the corresponding user groups.

3.1.2 Setup 2 (All samples). This setup also contains two analysis groups, A and B, each taking half of the population (i.e. $(n_0 + n_1 + n_2 + n_3)/2$). The metric mean and variance for groups A and B are the weighted metric mean and variance of the constituent user groups. As we only provide treatment to those who qualify for strategy 1 in group A, and likewise for group B with strategy 2, each user group will give different responses, e.g. for group A:

$$\mu_A = (n_0\mu_{C0} + n_1\mu_{I1} + n_2\mu_{C2} + n_3\mu_{I\phi})/(n_0 + n_1 + n_2 + n_3), \quad (9)$$

$$\sigma_A^2 = (n_0\sigma_{C0}^2 + n_1\sigma_{I1}^2 + n_2\sigma_{C2}^2 + n_3\sigma_{I\phi}^2)/(n_0 + n_1 + n_2 + n_3). \quad (10)$$

Substituting the above (and that for group B) into the definitions of actual effect size and MDE we have:

$$\Delta_{S2} = \frac{n_1(\mu_{C1} - \mu_{I1}) + n_2(\mu_{I2} - \mu_{C2}) + n_3(\mu_{I\psi} - \mu_{I_\phi})}{n_0 + n_1 + n_2 + n_3}, \quad (11)$$

$$\theta_{S2}^* = (z_{1-\alpha/2} - z_{1-\pi_{\min}}) \times \sqrt{\frac{2\Big(n_0(2\sigma_{C0}^2) + n_1(\sigma_{I1}^2 + \sigma_{C1}^2) + n_2(\sigma_{C2}^2 + \sigma_{I2}^2) + n_3(\sigma_{I\phi}^2 + \sigma_{I\psi}^2)\Big)}{(n_0 + n_1 + n_2 + n_3)^2}}. \quad (12)$$

3.1.3 Setup 3 (Qualified users only). The setup is very similar to Setup 2, with members from user group 0 excluded. This leads to both analysis groups having $(n_1 + n_2 + n_3)/2$ users. The absence of group 0 users means they are not featured in the weighted metric mean and variance of the two analysis groups, e.g. for group A:

$$\mu_A = \frac{n_1\mu_{I1} + n_2\mu_{C2} + n_3\mu_{I\phi}}{n_1 + n_2 + n_3}, \; \sigma_A^2 = \frac{n_1\sigma_{I1}^2 + n_2\sigma_{C2}^2 + n_3\sigma_{I\phi}^2}{n_1 + n_2 + n_3}. \quad (13)$$

This leads to the following actual effect size and MDE for Setup 3:

$$\Delta_{S3} = \frac{n_1(\mu_{C1} - \mu_{I1}) + n_2(\mu_{I2} - \mu_{C2}) + n_3(\mu_{I\psi} - \mu_{I_\phi})}{n_1 + n_2 + n_3}, \quad (14)$$

$$\theta_{S3}^* = (z_{1-\alpha/2} - z_{1-\pi_{\min}})\sqrt{\frac{2\Big(n_1(\sigma_{I1}^2 + \sigma_{C1}^2) + n_2(\sigma_{C2}^2 + \sigma_{I2}^2) + n_3(\sigma_{I\phi}^2 + \sigma_{I\psi}^2)\Big)}{(n_1 + n_2 + n_3)^2}}. \quad (15)$$

3.1.4 Setup 4 (Dual control). The setup is the odd one out as it has four analysis groups. Two of the analysis groups ($A1$ and $A2$) are drawn from those who qualified into strategy 1 and are allocated into the first randomization group, and the other two ($B1$ and $B2$) are drawn from those who are qualified into strategy 2 and are allocated into the second randomization group:

$$n_{A1} = n_{A2} = (n_1 + n_3)/4, \quad n_{B1} = n_{B2} = (n_2 + n_3)/4. \quad (16)$$

The metric mean and variance for group $A1$ are:

$$\mu_{A1} = \frac{n_1\mu_{C1} + n_3\mu_{C3}}{n_1 + n_3}, \; \sigma_{A1}^2 = \frac{n_1\sigma_{C1}^2 + n_3\sigma_{C3}^2}{n_1 + n_3}. \quad (17)$$

As the setup takes the difference of differences in the metric (i.e. the difference between groups $B2$ and $B1$, and the difference between groups $A2$ and $A1$), the actual effect size is as follows:

$$\begin{aligned}\Delta_{S4} &= (\mu_{B2} - \mu_{B1}) - (\mu_{A2} - \mu_{A1}) \\ &= \frac{n_2(\mu_{I2} - \mu_{C2}) + n_3(\mu_{I\psi} - \mu_{C3})}{n_2 + n_3} - \frac{n_2(\mu_{I1} - \mu_{C1}) + n_3(\mu_{I\phi} - \mu_{C3})}{n_1 + n_3}. \quad (18)\end{aligned}$$

The MDE for Setup 4 is similar to that specified in RHS of Ineq. (5), albeit with more groups:

$$\begin{aligned}\theta_{S4}^* &= (z_{1-\alpha/2} - z_{1-\pi_{\min}})\sqrt{\sigma_{A1}^2/n_{A1} + \sigma_{A2}^2/n_{A2} + \sigma_{B1}^2/n_{B1} + \sigma_{B2}^2/n_{B2}} \\ &= 2 \cdot (z_{1-\alpha/2} - z_{1-\pi_{\min}}) \times \\ &\sqrt{\frac{n_1(\sigma_{C1}^2 + \sigma_{I1}^2) + n_3(\sigma_{C3}^2 + \sigma_{I\phi}^2)}{(n_1 + n_3)^2} + \frac{n_2(\sigma_{C2}^2 + \sigma_{I2}^2) + n_3(\sigma_{C3}^2 + \sigma_{I\psi}^2)}{(n_2 + n_3)^2}}. \quad (19)\end{aligned}$$

3.2 Is dilution always bad?

The use of responses from users who do not qualify for any of the strategies we are comparing, an act known as metric dilution, has stirred countless debates in experimentation teams. On one hand, responses from these users make any treatment effect less pronounced by contributing exactly zero; on the other hand, it might be necessary as one does not know who actually qualify [5], or it might be desirable as they can be leveraged to reduce the variance of the treatment effect estimator [2].

Here, we are interested in whether we should engage in the act of dilution given the assumed user responses prior to an experiment. This can be clarified by understanding the conditions where Setup 3 would emerge superior (as defined in Sec. 2.3) to Setup 2. By inspecting Eqs. (11) and (14), it is clear that $\Delta_{S3} > \Delta_{S2}$ if $n_0 > 0$. Thus, Setup 3 is superior to Setup 2 under the first criterion if $\theta_{S3}^* < \theta_{S2}^*$, which is the case if σ_{C0}^2, the metric variance of users who qualify for neither strategies, is large. This can be shown by substituting Eqs. (12) and (15) into the θ-inequality and rearranging the terms to obtain:

$$\frac{\Big(n_1(\sigma_{I1}^2 + \sigma_{C1}^2) + n_2(\sigma_{C2}^2 + \sigma_{I2}^2) + n_3(\sigma_{I\phi}^2 + \sigma_{I\psi}^2)\Big) \cdot (n_0 + 2n_1 + 2n_2 + 2n_3)}{2(n_1 + n_2 + n_3)^2} < \sigma_{C0}^2. \quad (20)$$

If we assume the metric variance does not vary much for users who qualified for at least one strategy, i.e. $\sigma_{I1}^2 \approx \sigma_{C1}^2 \approx \cdots \approx \sigma_{I\psi}^2 \approx \sigma_S^2$, Ineq. (20) can then be simplified as

$$\sigma_S^2\left(\frac{n_0}{n_1 + n_2 + n_3} + 2\right) < \sigma_{C0}^2, \quad (21)$$

where it can be used to quickly determine if one should consider dilution at all.

If Ineq. (20) is not true (i.e. $\theta_{S3}^* \geq \theta_{S2}^*$), we should then consider when the second criterion (i.e. $\Delta_{S3} - \Delta_{S2} > \theta_{S3}^* - \theta_{S2}^*$) is met. Writing

$$\begin{aligned}\eta &= n_1(\mu_{C1} - \mu_{I1}) + n_2(\mu_{I2} - \mu_{C2}) + n_3(\mu_{I\psi} - \mu_{I\phi}), \\ \xi &= n_1(\sigma_{C1}^2 + \sigma_{I1}^2) + n_2(\sigma_{I2}^2 + \sigma_{C2}^2) + n_3(\sigma_{I\psi}^2 + \sigma_{I\phi}^2), \text{and} \\ z &= z_{1-\alpha/2} - z_{1-\pi_{\min}},\end{aligned}$$

we can substitute Eqs. (11), (12), (14), (15) into the inequality and rearrange to obtain

$$\frac{n_1 + n_2 + n_3}{n_0}\sqrt{2n_0\sigma_{C0}^2 + \xi} > \frac{n_0 + n_1 + n_2 + n_3}{n_0}\sqrt{\xi} - \frac{\eta}{\sqrt{2}z}. \quad (22)$$

As the LHS of Ineq. (22) is always positive, Setup 3 is superior if the RHS ≤ 0. Noting

$$\Delta_{S3} = \eta/(n_1 + n_2 + n_3) \text{ and } \theta_{S3}^* = \sqrt{2} \cdot z \cdot \sqrt{\xi}/(n_1 + n_2 + n_3),$$

the trivial case is satisfied if $(n_0 + n_1 + n_2 + n_3)/(n_0) \cdot \theta_{S3}^* \leq \Delta_{S3}$.

If the RHS of Ineq. (22) is positive, we can safely square both sides and use the identities for Δ_{S3} and θ^*_{S3} to get

$$\frac{2\sigma^2_{C0}}{n_0} > \frac{\left[\left(\theta^*_{S3} - \Delta_{S3} + \frac{n_1+n_2+n_3}{n_0}\theta^*_{S3}\right)^2 - \left(\frac{n_1+n_2+n_3}{n_0}\theta^*_{S3}\right)^2\right]}{2z^2}. \quad (23)$$

As the LHS is always positive, the second criterion is met if

$$\theta^*_{S3} \leq \Delta_{S3}. \quad (24)$$

Note this is a weaker, and thus more easily satisfiable condition than that introduced in the previous paragraph. This suggests an experiment setup is always superior to an diluted alternative if the experiment is already adequately powered. Introducing any dilution will simply make things worse.

Failing the condition in Ineq. (24), we can always fall back to Ineq. (23). While the inequality operates in squared space, it is essentially comparing the standard error of user group 0 (LHS) — those who qualify for neither strategies — to the gap between the minimum detectable and actual effects ($\theta^*_{S3} - \Delta_{S3}$). The gap can be interpreted as the existing noise level, thus a higher standard error means mixing in group 0 users will introduce extra noise, and one is better off without them. Conversely, a smaller standard error means group 0 users can lower the noise level, i.e. stabilize the metric fluctuation, and one should take advantage of them.

To summarize, diluting a personalization strategies experiment setup is *not* helpful if (i) users who do not qualify for any strategies have a large metric variance (Ineq. (20)) or (ii) the experiment is already adequately powered (Ineq. (24)). It could help if the experiment has not gained sufficient power yet and users who do not qualify for any strategy provide low-variance responses, such that they exhibit stabilizing effects when included into the analysis (complement of Ineq. (23)).

3.3 When is a dual-control more effective?

Often when advertisers compare two personalization strategies, the question on whether to use a dual control/multi-cell design comes up. Proponents of such approach celebrate its ability to tell a story by making the incrementality of an individual strategy available, while opponents voice concerns on the complexity in setting up the design. Here we are interested if Setup 4 (dual control) is superior to Setup 3 (a simple A/B test) from a power/detectable effect perspective, and if so, under what circumstances.

We first observe $\theta^*_{S4} > \theta^*_{S3}$ is always true, and hence a dual control setup will never be superior to a simpler setup under the first criterion. This can be verified by substituting in Eqs. (19) and (15) and rearranging the terms to show the inequality is equivalent to

$$\begin{aligned}&2\Big(\frac{n_1}{(n_1+n_3)^2}(\sigma^2_{C1}+\sigma^2_{I1}) + \frac{n_2}{(n_2+n_3)^2}(\sigma^2_{C2}+\sigma^2_{I2}) + \frac{n_3}{(n_1+n_3)^2}\sigma^2_{I\phi}+\\ &\quad\frac{n_3}{(n_2+n_3)^2}\sigma^2_{I\psi} + \left(\frac{n_3}{(n_1+n_3)^2}+\frac{n_3}{(n_2+n_3)^2}\right)\sigma^2_{C3}\Big) >\\ &\frac{n_1}{(n_1+n_2+n_3)^2}(\sigma^2_{C1}+\sigma^2_{I1}) + \frac{n_2}{(n_1+n_2+n_3)^2}(\sigma^2_{C2}+\sigma^2_{I2})+\\ &\quad\frac{n_3}{(n_1+n_2+n_3)^2}\sigma^2_{I\phi} + \frac{n_3}{(n_1+n_2+n_3)^2}\sigma^2_{I\psi},\end{aligned} \quad (25)$$

which is always true given the ns are non-negative and the σ^2s are positive: not only the coefficients of the σ^2-terms are larger on the LHS than their RHS counterparts, the LHS also carries an extra σ^2_{C3} term with non-negative coefficient and a factor of two.

Moving on to the second evaluation criterion, we recall that Setup 4 is superior if $\Delta_{S4} - \Delta_{S3} > \theta^*_{S4} - \theta^*_{S3}$, otherwise Setup 3 is superior under the same criterion. The full flexibility of the model can be seen by substituting Eqs. (14), (15), (18), and (19) into the inequality and rearrange to obtain

$$\begin{aligned}&\frac{n_1\frac{n_2(\mu_{I2}-\mu_{C2})+n_3(\mu_{I\psi}-\mu_{C3})}{n_2+n_3} + n_2\frac{n_1(\mu_{I1}-\mu_{C1})+n_3(\mu_{I\phi}-\mu_{C3})}{n_1+n_3}}{\sqrt{n_1(\sigma^2_{C1}+\sigma^2_{I1}) + n_2(\sigma^2_{C2}+\sigma^2_{I2}) + n_3(\sigma^2_{I\phi}+\sigma^2_{I\psi})}} > \quad (26)\\ &\sqrt{2}z\left[\sqrt{2\cdot\frac{\begin{array}{c}(1+\frac{n_2}{n_1+n_3})^2[n_1(\sigma^2_{C1}+\sigma^2_{I1})+n_3(\sigma^2_{C3}+\sigma^2_{I\phi})]+\\ (1+\frac{n_1}{n_2+n_3})^2[n_2(\sigma^2_{C2}+\sigma^2_{I2})+n_3(\sigma^2_{C3}+\sigma^2_{I\psi})]\end{array}}{n_1(\sigma^2_{C1}+\sigma^2_{I1})+n_2(\sigma^2_{C2}+\sigma^2_{I2})+n_3(\sigma^2_{I\phi}+\sigma^2_{I\psi})}} - 1\right],\end{aligned}$$

where $z = z_{1-\alpha/2} - z_{1-\pi_{\min}}$.

A key observation from inspecting Ineq. (26) is that the LHS of the inequality scales along $O(\sqrt{n})$, while the RHS remains a constant. This leads to the insight that Setup 4 is more likely to be superior if the ns are large. Here we assume the ratio $n_1 : n_2 : n_3$ remains unchanged when we scale the number of samples, an assumption that generally holds when an organization increases their reach while maintaining their user mix. It is worth pointing out that our claim is stronger than that in previous work — we have shown that having a large user base not only fulfills the requirement of running a dual control experiment as described in [5], it also makes a dual control experiment a better setup than its simpler counterparts in terms of apparent and detectable effect sizes.

The scaling relationship can be seen more clearly if we apply some simplifying assumptions to the σ^2- and n-terms. If we assume the metric variances are similar across user groups (i.e. $\sigma^2_{C1} \approx \sigma^2_{I1} \approx \cdots \approx \sigma^2_{I\psi} \approx \sigma^2_S$), the RHS of Ineq. (26) becomes

$$\sqrt{2}z\left[\sqrt{\frac{n_1+n_2+n_3}{n_1+n_3} + \frac{n_1+n_2+n_3}{n_2+n_3}} - 1\right], \quad (27)$$

which remains a constant if the ratio $n_1 : n_2 : n_3$ remains unchanged. If we assume the number of users in groups 1, 2, 3 are similar (i.e. $n_1 = n_2 = n_3 = n$), the LHS of Ineq. (26) becomes

$$\frac{\sqrt{n}\big((\mu_{I2}-\mu_{C2}) - (\mu_{I1}-\mu_{C1}) + \mu_{I\psi} - \mu_{I\phi}\big)}{2\sqrt{\sigma^2_{C1}+\sigma^2_{I1}+\sigma^2_{C2}+\sigma^2_{I2}+\sigma^2_{I\phi}+\sigma^2_{I\psi}}}, \quad (28)$$

which clearly scales along $O(\sqrt{n})$.

We conclude the section by providing an indication on what a large n may look like, if we assume both the metric variances and the number of users are are similar across user groups, we can rearrange Ineq. (26) to make n the subject:

$$n > \left(2\sqrt{12}\left(\sqrt{6}-1\right)z\right)^2\frac{\sigma^2_S}{\Delta^2}, \quad (29)$$

where $\Delta = (\mu_{I2} - \mu_{C2}) - (\mu_{I1} - \mu_{C1}) + \mu_{I\psi} - \mu_{I\phi}$ is the effect size. With a 5% significance level and 80% power, the first coefficient amounts to around 791, which is roughly 50 times the coefficient one would use to determine the sample size of a simple A/B test [6]. This suggests a dual control setup is perhaps a luxury accessible only to the largest advertising platforms and their top advertisers. For example, consider an experiment to optimize conversion rate where the baselines attain 20% (hence having a metric variance of

	Actual effect size	Minimum detectable effect
Setup 1	1049/1099 (95.45%)	66/81 (81.48%)
Setup 2	853/999 (85.38%)	87/106 (82.08%)
Setup 3	922/1099 (83.89%)	93/116 (80.18%)
Setup 4	240/333 (72.07%)	149/185 (80.54%)

Table 2: Number of evaluations where the theoretical value of the quantities (columns) falls between the 95% bootstrap confidence interval for each experiment setup (rows). See Section 4 for a detailed description on the evaluations.

$0.2(1 - 0.2) = 0.16$). If there is a 2.5% relative (i.e. 0.5% absolute) effect between the competing strategies, the dual control setup will only be superior if there are > 5M users in each user group.

4 EXPERIMENTS

Having performed theoretical calculations for the actual and detectable effects and conditions where an experiment setup is superior to another, here we verify those calculations using simulation results. We focus on the results presented in Section 3.1, as the rest of the results presented followed from those calculations.

In each experiment setup evaluation, we randomly select the value of the parameters (i.e. the μs, σ^2s, and ns), and take 1,000 actual effect samples, each by (i) sampling the responses from the user groups under the specified parameters, (ii) computing the mean for the analysis groups, and (iii) taking the difference of the means.

We also take 100 MDE samples in separate evaluations, each by (i) sampling a critical value under null hypothesis; (ii) computing the test power under a large number of possible effect sizes, each using the critical value and sampled metric means under the alternate hypothesis; and (iii) searching the effect size space for the value that gives the predefined power. As the power vs. effect size curve is noisy given the use of simulated power samples, we use the bisection algorithm provided by the `noisyopt` package to perform the search. The algorithm dynamically adjusts the number of samples taken from the same point on the curve to ensure the noise does not send us down the wrong search space.

We expect the mean of the sampled actual effect and MDE to match the theoretical value. To verify this, we perform 1,000 bootstrap resamplings on the samples obtained above to obtain an empirical bootstrap distribution of the sample mean in each evaluation. The 95% bootstrap resampling confidence interval (BRCI) should then contain the theoretical mean 95% of the times. The histogram of the percentile rank of the theoretical quantity in relation to the bootstrap samples across multiple evaluations should also follow a uniform distribution [8].

The result is shown in Table 2. One can observe that there are more evaluations having their theoretical quantity lying outside than the BRCI than expected. Upon further investigation, we observed a characteristic ∪-shape from the histograms of the percentile ranks for the actual effects. This suggests the bootstrap samples may be under-dispersed but otherwise centered on the theoretical quantities.

We also observed the histograms for MDEs curving upward to the right, this suggests that the theoretical value is a slight overestimate (of $<$ 1% to the bootstrap mean in all cases). We believe this is likely due to a small bias in the bisection algorithm. The algorithm tests if the mean of the power samples is less than the target power to decide which half of the search space to continue along. Given we can bisect up to 10 times in each evaluation, it is likely to see a false positive even when we set the significance level for individual comparisons to 1%. This leads to the algorithm favoring a smaller MDE sample. Having that said, since we have tested for a wide range of parameters and the overall bias is small, we are satisfied with the theoretical quantities for experiment design purposes.

5 CONCLUSION

We have addressed the problem of comparing experiment designs for personalization strategies by presenting an evaluation framework that allows experimenters to evaluate which experiment setup should be adopted given the situation. The flexible framework can be easily extended to compare setups that compare more than two strategies by adding more user groups (i.e. new sets to the Venn diagram in Fig. 1). A new setup can also be incorporated quickly as it is essentially a different weighting of user group-scenario combinations shown in Table 1. The framework also allows the development of simple rule of thumbs such as:

(i) Metric dilution should never be employed if the experiment already has sufficient power; though it can be useful if the experiment is under-powered and the non-qualifying users provide a "stabilizing effect"; and
(ii) A dual control setup is superior to simpler setups only if one has access to the user base of the largest organizations.

We have validated the theoretical results via simulations, and made the code available[2] so that practitioners can benefit from the results immediately when designing their upcoming experiments.

ACKNOWLEDGMENTS

The work is partially funded by the EPSRC CDT in Modern Statistics and Statistical Machine Learning at Imperial and Oxford (StatML.IO) and ASOS.com. The authors thank the anonymous reviewers for providing many improvements to the original manuscript.

REFERENCES

[1] Brooke Bengier and Amanda Knupp. [n.d.]. Selling More Stuff: The What, Why & How of Incrementality Testing. https://www.quantcast.com/blog/selling-more-stuff-the-what-why-how-of-incrementality-testing/. Blog post.

[2] Alex Deng and Victor Hu. 2015. Diluted Treatment Effect Estimation for Trigger Analysis in Online Controlled Experiments. In *WSDM '15* (Shanghai, China). 349–358. https://doi.org/10.1145/2684822.2685307

[3] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 2013. Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-experiment Data. In *WSDM '13* (Rome, Italy). 123–132. https://doi.org/10.1145/2433396.2433413

[4] Pavel Dmitriev, Somit Gupta, Dong Woo Kim, and Garnet Vaz. 2017. A Dirty Dozen: Twelve Common Metric Interpretation Pitfalls in Online Controlled Experiments. In *KDD '17* (Halifax, NS, Canada). 1427–1436.

[5] C. H. Bryan Liu, Elaine M. Bettaney, and Benjamin Paul Chamberlain. 2018. Designing Experiments to Measure Incrementality on Facebook. arXiv:1806.02588. [stat.ME] 2018 AdKDD & TargetAd Workshop.

[6] Evan Miller. 2010. How Not To Run an A/B Test. https://www.evanmiller.org/how-not-to-run-an-ab-test.html. Blog post.

[7] Alexey Poyarkov, Alexey Drutsa, Andrey Khalyavin, Gleb Gusev, and Pavel Serdyukov. 2016. Boosted Decision Tree Regression Adjustment for Variance Reduction in Online Controlled Experiments. In *KDD '16* (San Francisco, California, USA). 235–244. https://doi.org/10.1145/2939672.2939688

[8] Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. 2018. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. arXiv:1804.06788 [stat.ME]