Balancing Precision and Retention in Experimental Design

May 10, 2023

Abstract

In experimental social science, precise treatment effect estimation is of utmost importance, and researchers can make design choices to increase precision. Specifically, block-randomized and pre-post designs are promoted as effective means to increase precision. However, implementing these designs requires pre-treatment covariates, and collecting this information may decrease sample sizes, which in and of itself harms precision. Therefore, despite the literature's recommendation to use block-randomized and pre-post designs, it remains unclear whether these designs increase precision in applied settings. In this article, we present guidelines to assist researchers in navigating these design decisions. Using replication and simulated data, we demonstrate a counterintuitive result: precision gains from block-randomized or pre-post designs can withstand significant sample loss that may arise during implementation. Our findings underscore the importance of incorporating researchers' practical concerns into existing experimental design advice.

Absract word count: 135
Manuscript word count: 10,583

1 Introduction

Research design for randomized experiments is an area of active innovation in the social sciences (Druckman and Green 2021). With new tools to simulate design choices (Blair et al. 2019) and new norms like preregistration (Ofosu and Posner 2021), researchers are pushed to consider the properties of their research design before collecting data. Amid growing concerns over the lack of statistical power in most of quantitative political science (Arel-Bundock et al. 2022), one important property is the precision of the procedure used to estimate treatment effects with experimental data. Researchers have only one chance to conduct randomization, collect data, and generate an estimate of the average treatment effect (ATE). The stakes are high, so decreasing the statistical variability of the research design is key to detecting non-zero treatment effects when they exist.

Fortunately, researchers can consider multiple practices to improve precision at the research design stage. Some strategies are increasing sample size if resources permit, even if only in the control group (A. S. Gerber, Green, and Larimer 2008), using placebo conditions instead of pure controls to account for features of a treatment that are not relevant (Broockman, Kalla, and Sekhon 2017), choosing the right balance of abstraction and detail when crafting survey vignettes (Brutger et al. 2020), incentivizing survey attention (Berinsky, Margolis, and Sances 2014; Kane, Velez, and Barabas 2023), or using an index instead of a single outcome variable to reduce measurement error (Broockman, Kalla, and Sekhon 2017). All these strategies and more can increase precision, and thus make an experiment more likely to recover an estimate closer to the true ATE.

In this article, we revisit two precision-improving research designs that are promoted as particularly effective. We consider (1) block randomization and (2) pre-post designs. With block randomization, researchers create subgroups of units they expect to respond similarly to treatment. Then, randomization occurs separately within these groups, rather than across the entire sample. This improves precision by reducing variation in potential outcomes within

blocks (Imai, King, and Stuart 2008). Pre-post designs adjust for a pre-treatment measure of the outcome, which can improve precision by controlling for a major source of variability (Clifford, Sheagley, and Piston 2021).

We assess block randomization and pre-post designs for four main reasons. First, we focus on these designs because the literature explicitly recommends using them as particularly effective ways to increase precision. Research shows that blocking is unlikely to hurt (Imai, King, and Stuart 2008; Pashley and Miratrix 2021a) and can greatly improve precision in applied settings (Moore 2012), hence the slogan "block what you can, randomize what you cannot" (Box et al. 1978, 103). Regarding repeated measures designs, recent work recommends researchers to implement this design "whenever possible" to improve precision (Clifford, Sheagley, and Piston 2021, 1062).

Second, and most importantly, we focus on these designs because of literature's promise of improved precision **assumes sample size is not affected** by the decision to implement these designs. Yet in practice, implementing these design may decrease sample size, offsetting promised precision gains, and making it unclear how researchers should navigate these decisions. Specifically, there are two ways these design choices may negatively affect sample size. *Explicit sample loss* happens when units or subjects recruited for a study drop from the experiment under an alternative design when they would not under the standard design. For example, subjects recruited in a pre-treatment survey wave may not be available again for the second wave containing the experimental manipulation. *Implicit sample loss* happens when investing in an alternative design forces the researcher to settle with a smaller sample size for budgeting reasons. For example, the choice to conduct both pre-treatment and post-treatment surveys could lead a researcher to settle for a smaller sample than if a researcher devoted the entire budget to only measuring outcomes after delivering treatment. A design choice's promise of increased precision should be questioned if sample size is adversely affected by it, either explicitly or implicitly. This is an important practical concern when implementing

these designs that we build on the existing literature to address.

Third, despite the decisive advice in the literature that block randomized and pre-post designs provide precision gains, they are not widely implemented in experimental political science. Since its inception in 2014 up to the time of writing, only 21 out of 245 ($\approx 9\%$) articles in the *Journal of Experimental Political Science* mention block randomization.[1] For repeated measures, Clifford, Sheagley, and Piston (2021) recently sampled articles from experiment-friendly political science journals and found that only 18% deviate from measuring only post-treatment outcomes.[2]

Fourth, we focus on block randomization and pre-post designs because they represent a broader class of design choices that require researchers to decide whether it is worth it to collect pre-treatment information about covariates or outcomes, and if so, how to use it. Techniques like block randomization increase precision via the randomization procedure, whereas pre-post designs reflect strategies that increase precision by reducing noise in measured outcomes. Therefore, considering these two design choices allows us to assist researchers in choosing not just whether to measure pre-treatment information, but also what to do with it.

For these reasons, we develop guidelines to navigate the choice to implement block randomized and pre-post designs when a researcher risks sample loss from these choices. We investigate the problem from three perspectives to inform our guidelines. First, we make the competing components of precision clear, taking a broad view of how sample size can be affected both explicitly and implicitly from alternative designs. Second, we revisit the research design of a survey experiment on the effect of social media comments on belief in misinformation (Anspach and Carlson 2020) to illustrate how a researcher can entertain alternative designs before conducting an experiment. Third, we use simulations to show how applied researchers

[1] We focus on this journal since it publishes experiments almost exclusively across a variety of subfields. 33 articles mention "blocks", "blocking", or related terms. We exclude the articles discussing reporting standards, as well as articles that use the terms to refer to "blocks" in contexts other than block randomization.

[2] We are working on extending these figures to produce a more comprehensive analysis of general interest journals in Political Science and will update accordingly prior to publication.

can consider these experimental designs at the design-stage. To facilitate the translation between our findings and future applications, Appendix A presents a flowchart with steps to consider how alternative designs balance precision amid potential sample loss.

We echo the advice in the literature that block randomized and pre-post designs improve precision, but draw attention to the fact this requires sample size not be affected. The more complicated scenario arises when sample loss might occur. Critically, our simulations show that highly predictive blocking covariate(s) and pre-treatment outcome measure(s) can produce precision gains that *withstand non-negligible sample loss*. In other words, perhaps running counter to what a researcher would expect, the optimal experimental design may involve sacrificing sample size in order to implement block randomization or pre-post measurement. We also provide cautionary advice. We show that these designs may inadvertently harm precision if sample loss is likely to occur from their implementation and the incorporated pre-treatment information is not strongly predictive of the outcomes.

This article contributes to a growing literature highlighting the merits and applicability of alternative experimental designs. For example, Clifford, Sheagley, and Piston (2021) show that researchers can implement a repeated measures design within the same survey without worrying about altering treatment effects, Pashley and Miratrix (2021a) show that block-randomization is guaranteed to improve statistical precision, or at least not hurt it, in experiments with an equal proportion of treated units across blocks. By developing guidelines to determine whether the investment on alternative designs is worth it, we further expand researchers' ability to implement the appropriate experiment across a broad range of applications. Moreover, our guidelines can also assist resource-constrained researchers to maximize statistical precision under limited budgets.

2 Designs to Improve Precision and Sample Loss

The most common experimental design implemented in the social sciences has two defining features. First, it assigns treatments using simple or complete randomization (see Bowers and Leavitt 2020 for details). Second, it measures outcomes only after administering treatments. We refer to a design using simple or complete randomization and post-treatment outcomes measurement only as the "standard" design.

Precision concerns motivate researchers to entertain alternative research designs.[3] Researchers can deviate from the standard design by choosing an alternative randomization procedure or an alternative timing of the outcome measurement. Table 1 maps the alternative designs that we consider across these two dimensions. In this article, we compare the standard design to three other designs highlighted in Table 1: block randomized, pre-post, and their combination. We discuss each in turn.

2.1 Block randomization

First, consider block randomization. Block randomization (or blocking) randomly assigns treatment within subgroups of units that the researcher expects will respond similarly to the experimental interventions. This randomization procedure is advantageous because it creates mini-experiments where the treatment and control groups' potential outcomes are as similar as possible. Block randomized designs can greatly improve precision in social science applications (e.g., Moore 2012) and thus are highly recommended in the literature (e.g. King et al. 2007; Moore and Moore 2013; Imai, King, and Stuart 2008; Pashley and Miratrix 2021b, 2021a). In fact, Imai, King, and Stuart (2008) advise that when feasible, "blocking on potentially confounding covariates should always be used" (493).

[3]Other motivations, like features of the research context, may also lead researchers to consider deviations from the standard design. For example, coordinating field experiments across two distant sites may turn complete randomization across sites challenging, forcing the researcher to conduct independent experiments in each site, which would be equivalent to block randomization. We focus on cases where researchers can reasonably entertain the research design choices in Table 1.

Table 1: Alternatives to the standard experimental design

	Outcome measurement	
	Post-only	Pre-post
Randomization		
Complete	Standard	Pre-post
Block	Block randomized	Block randomized & pre-post measures

2.2 Pre-post design

The second dimension of design choices we consider is the timing of outcome measurement. The majority of experiments in political science measure outcomes only post-treatment and compare observed outcomes across treatment and control groups to estimate treatment effects (Clifford, Sheagley, and Piston 2021). Precision can improve if pre-treatment measures of the outcomes are also collected and used in one of two ways. First, pre-treatment measures can be used to rescale the outcome as the difference between the two measures. Second, pre-treatment outcome measures can also be used on the right hand side of a regression model of treatment effects as a form of covariate adjustment. A pre-treatment measure of the outcome is often the best predictor of a unit's observed outcome, so controlling for this one piece of information can greatly improve precision in estimated treatment effects.

Collecting pre-treatment outcomes is considered good practice in field experiments, especially those using surveys to measure outcomes (Broockman, Kalla, and Sekhon 2017). However, they are less common in one-wave survey experiments, since their inclusion can lead to priming effects that alter the potential outcomes in undesirable ways. To address this problem, Clifford, Sheagley, and Piston (2021) replicate several survey experiments, including pre-treatment outcome measurements, to show that priming effects are rarely a problem. They also consider alternative versions of the repeated measures design for settings where priming is a concern. For example, a quasi pre-post design measures a proxy of the outcome of interest before treatment and the actual outcome afterwards to minimize priming effects.

2.3 Explicit and implicit sample loss

While these design choices have statistical benefits, we draw attention to an important practical concern that often arises when researchers consider implementing them over the standard design— a study may lose sample size as a result. Therefore, researchers may be cautious about using block randomized or pre-post designs. The literature's recommendation to use these designs assumes sample size is unaffected by the decision to implement them. However, in practice, researchers often run into contexts where that is unlikely to be the case, leaving the conditions under which it is advantageous to implement these designs unclear. In this section, we outline how sample size could be attenuated either explicitly or implicitly. We do so to make these consequences a part of decision-making process when considering alternative designs to improve precision.

First, we refer to "explicit sample loss" as circumstances when the sample is already defined and units who would finish the experiment under the standard design do not finish it under an alternative design. For example, this could occur if the block randomization procedure discarded units that would have been randomized to treatment under complete randomization. This type of sample loss could also occur if the researcher adds many covariates to a pre-treatment survey for blocking or repeated measures purposes, increasing survey fatigue. As a result, more units might provide noisy or missing data or even drop from the survey than would under the standard design where these additional covariates are not asked pre-treatment.

We also draw attention to the scenario where sample loss occurs implicitly. We refer to "implicit sample loss" as loss happening when investing in an alternative design leads the researcher to settle with a smaller sample size before the study is even fielded. This means implicit sample loss is not something one can gather from looking at an experiment's raw data. For example, with a set budget, a researcher may settle with a smaller sample size in order to ask more questions in a pre-treatment survey. Because her budget would have afforded her more units if she only asked questions post-treatment according to the standard

design, we call this implicit sample loss from the alternative design.

Because circumstances and decisions that lead to implicit sample loss are not usually included in published research, we provide two toy examples to understand how implicit sample loss occurs at the design stage of a study. First, imagine a researcher wishes to conduct a survey experiment with Prolific. Using this platform, a five minute survey with a non-representative sample of 1,000 respondents costs USD$1,173. Adding four extra pre-treatment questions that require two more minutes to complete for the average participant increases the cost to $1,640. To keep the extra questions and stay within budget, the researcher would need to reduce the sample size to about 720 respondents.[4] The four additional pre-treatment questions would allow the researcher to use a block randomized or pre-post design, but is 72% of a researcher's potential sample a good trade off?

Second, consider a field experiment needing to conduct an additional survey wave to collect pre-treatment covariates. This is an extreme case that would imply, all else constant, the cost of data collection doubles (i.e., administering two surveys instead of one). With a fixed budget, this translates to retaining half of the sample that a standard design experiment would enjoy due to implicit sample loss. Again, the additional pre-treatment information could have large precision-increasing effects, but is it worth it to collect this information if the researcher can then only afford half as many subjects? In this article, we outline how a researcher can approach these questions.

2.4 Sample loss and bias

Depending on its nature, sample loss can also introduce bias in treatment effect estimates. For example, this would be the case if the factors that cause explicit sample loss correlate with pre-treatment outcomes or blocking covariates. This is not the primary focus of this article

[4]This follows from the cost calculator in https://www.prolific.co/old/pricing for academic/non-profit purposes at the default hourly rate of USD$10.54 per respondent as of February 21, 2023. We choose this platform for the example exclusively because of the easy access to the cost calculator.

since the implications of sample loss for statistical precision persist even when sample loss does not correlate with potential outcomes or key covariates. In other words, uncorrelated, or random, sample loss is sufficient to illustrate the challenges in optimizing precision when choosing among alternative experimental designs. We show via simulation in Appendix E that introducing bias adds more nuance to the act of balancing precision and retention, but the main conclusions of this article still stand. Furthermore, introducing bias as another consideration on top of precision and sample loss requires statements about the nature and direction of bias that researchers can only inform with domain expertise.

Correlated attrition may introduce bias in the case of explicit sample loss. For example, sample loss could be related to covariates that drive treatment effect heterogeneity, like if people attrit from online experiments due to lack of digital literacy (Guess and Munger 2020). However, previous research has already outlined strategies to address and mitigate this form of attrition that apply to the designs discussed in this article as much as they apply to any experiment (A. Gerber et al. 2014; Coppock et al. 2017; Lo, Renshon, and Bassan-Nygate 2023).

Finally, correlated attrition is largely not a concern with implicit sample loss. If the experiment obtains a smaller random sample from the population for budget reasons, this would not introduce bias. Likewise, if a field experiment considers whether it has the budget to sample households from one state or two (e.g. Nickerson 2008), this decision concerns the population and estimand, not bias.

3 Balancing Precision and Retention Under Alternative Designs

In this section, we illustrate the tension between precision and retention by describing the standard experimental research design. We then demonstrate how the alternative designs

in Table 1 also facilitate unbiased estimators of the ATE and discuss how these alternative designs improve precision.

3.1 The standard experiment

Consider an experiment in a sample of N units indexed by $i = \{1, 2, 3, ..., N\}$. For simplicity, consider a binary treatment so that $Z_i = \{0, 1\}$ denotes unit i's treatment assignment. Using the Neyman–Rubin potential outcomes framework, assume two potential outcomes, one if a unit receives treatment ($Y_i(1)$) and one if the unit receives the control ($Y_i(0)$). In addition to assuming the potential outcomes satisfy SUTVA and excludability, we also assume treatment is randomly assigned.

The first defining feature of the standard experimental design pertains to the random assignment, which could be either complete or simple randomization. With a sufficiently large sample size, both randomization procedures yield equivalent treatment assignments in expectation, so we focus on complete randomization for the sake of exposition (see Bowers and Leavitt 2020 for details). With a binary treatment, complete randomization randomly permutes N units and assigns the first m units to treatment and the remaining $N - m$ to control. Thus, the vector of random treatment assignments $\boldsymbol{Z} = \{Z_1, ..., Z_N\}^\top$ contains a fixed number of m units assigned to treatment and $N - m$ assigned to control. Usually $m = N/2$, but the only requirement is that the m units are selected at random.

The second defining feature of the standard experiment is that it only measures outcomes after administering treatments. Unit i's potential outcomes relate to its observed outcome Y_i using the following switching equation: $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$, and Y_i is observed after treatment. In this article, we are interested in the average treatment effect as our estimand, as it is the most common quantity of interest in social science applications: $ATE = E[Y_i(1) - Y_i(0)]$. We can obtain an unbiased estimate of the ATE by calculating the difference in the average observed outcome in the treatment and control groups: $\widehat{ATE} =$

$E[Y_i(1)|Z_i = 1] - E[Y_i(0)|Z_i = 0] = \left[\frac{1}{m}\sum_{i=1}^{m} Y_i\right] - \left[\frac{1}{N-m}\sum_{m+1}^{N} Y_i\right]$.

The true standard error of the difference in means estimator (A. S. Gerber and Green 2012, 57) under the standard design is

$$SE(\widehat{ATE}_{\text{Standard}}) = \sqrt{\frac{\frac{m}{N-m}\text{Var}(Y_i(0)) + \frac{N-m}{m}\text{Var}(Y_i(1)) + 2\text{Cov}(Y_i(0), Y_i(1))}{N-1}}. \quad (1)$$

If we assume half of the participants are assigned to treatment and half to control ($m = N/2$) it simplifies to

$$SE(\widehat{ATE}_{\text{Standard}}) = \sqrt{\frac{\text{Var}(Y_i(0)) + \text{Var}(Y_i(1)) + 2\text{Cov}(Y_i(0), Y_i(1))}{N-1}}. \quad (2)$$

This formula represents the standard deviation of the distribution of all $\widehat{ATE}$'s given all possible random assignments.

The simplest alternative to improve precision would be to increase the sample size. Because of the factor $\frac{1}{\sqrt{N-1}}$ in $SE(\widehat{ATE}_{\text{Standard}})$, to cut the standard error in half under the standard design, a researcher would need four times the sample size. Increasing N enough to meaningfully increase precision is often not an option for researchers. In most applications this is cost prohibitive. Moreover, even if cost is not an issue, not all populations of interest can be increased to a trivially large sample size, as may be the case if conducting a survey experiment with a sample of Black Americans (Burge, Wamble, and Cuomo 2020) or white Evangelical Americans (Adida et al. 2022). Likewise, many field experiments cannot simply quadruple their sample size for logistical reasons, like recruiting enumerators or visiting locations, even if funds permitted.

When increasing N is not an option, we have discussed two design choices (block randomized and pre-post designs) that the literature promotes as effective ways to increase precision. To see how these designs increase precision, consider the standard error of the $\widehat{ATE}$ in Equation

2. Block randomized and pre-post designs achieve precision gains by reducing $Var(Y_i(0))$ and $Var(Y_i(1))$.[5]

However, any precision gains are called into question if block randomized and pre-post designs also inflict precision costs by reducing N. Put simply, as long as the numerator decreases *more* than the denominator decreases, or the variance in the potential outcomes decreases *more* than any resulting loss in sample size, the standard error will decrease in turn. This is how experimental design choices influence $SE(\widehat{ATE})$. We next explain in more detail how block randomized and pre-post designs accomplish this goal.

3.2 Block randomized experiment

One way to decrease $SE(\widehat{ATE})$ is to adjust the randomization procedure to block randomization. Block randomization requires the researcher collect pre-treatment covariates expected to correlate with potential outcomes. Then, the researcher groups observations into blocks or strata along these covariates, conducts randomization *within* each block, and combines results across blocks with a weighted difference-in-means estimator.

More formally, we now have B blocks and n_b units per block. In each block, we assign m_b units to treatment and $n_b - m_b$ units to control. The proportion of treated units per block does not need to be the same across blocks. Because randomization occurs within each block, we can consider each block as if we are conducting an independent experiment. The block-level ATE estimator is $\widehat{ATE}_b = E[Y_{ib}(1)|Z_{ib} = 1] - E[Y_{ib}(0)|Z_{ib} = 0]$. The most common estimator for the overall ATE combines estimates across blocks by weighting block-level $\widehat{ATE}_b$ depending on the size of the block. We call this estimator $\widehat{ATE}_{\text{Block}}$ to distinguish a block randomized design from the complete randomized design discussed above.

$$\widehat{ATE}_{\text{Block}} = \sum_{b=1}^{B} \frac{n_b}{N} \widehat{ATE}_b. \tag{3}$$

[5]We set aside the role of the covariance in potential outcomes by assuming it is held constant.

This estimator is unbiased but its precision depends on the correlation between potential outcomes and the proportion of treated units across blocks (Bowers, Diaz, and Grady 2022). Appendix C discusses alternative approaches to estimate overall ATEs in block randomized experiments.

We set aside the issue of variance estimation and consider how block randomization as a design choice affects the true standard error of the estimated ATE.[6] Like the $\widehat{ATE}_{\text{Block}}$, the true standard error is a weighted average of within-block standard errors (A. S. Gerber and Green 2012, 74):

$$SE(\widehat{ATE}_{\text{Block}}) = \sqrt{\sum_{b=1}^{B} \left(\frac{n_b}{N}\right)^2 SE^2(\widehat{ATE}_b)}. \tag{4}$$

Block randomized experiments yield more precise estimates than the standard design when the researcher creates blocks with covariates that correlate with potential outcomes. This is because the variance of the potential outcomes is smaller within each block (Imai, King, and Stuart 2008). Blocking can use a single covariate to stratify, like partisanship, or groups formed by overlapping key covariates, like partisanship and gender. The literature recommends blocking on all pre-treatment information available to a researcher using multivariate blocking procedures that collapse many variables into groups of comparable observations (Moore 2012). But, usually, researchers can only afford to block on discrete covariates, and these are collected at their own expense before administering treatments.[7]

The more $Var(Y_i(0))$ and $Var(Y_i(1))$ shrink *within each block*, the more the variance of the potential outcomes component of $SE(\widehat{ATE}_{\text{Block}})$ shrinks relative to the standard design. However, this statistical benefit only applies if sample size is not affected. By examining

[6] Optimal variance estimation depends on the composition of blocks. Imai (2008) discusses pair-matched experiments, a special case of block randomization in which $n_b = 2$ across blocks. Pashley and Miratrix (2021b) show how to estimate the variance in experiments that combine pair-matched and larger blocks.

[7] Sequential and online blocking algorithms alleviate both concerns by grouping observations as they join the experiment based on a pre-determined similarity metric (Higgins, Sävje, and Sekhon 2016; Moore and Moore 2013). But these usually require resources and sample sizes that may not be available in most applications.

Equation 2, we can see that if the denominator decreases as the numerator decreases, the positive effects of the block randomized design choice on precision are called into question.[8]

3.3 Pre-post design

Another way to decrease $SE(\widehat{ATE})$ is by measuring the outcome variable before treatment assignment in addition to the usual post-treatment measurement. We focus on what Clifford, Sheagley, and Piston (2021) refer to as the "between-subjects pre-post design," but we simply call it the "pre-post" design. The additional pre-treatment information gathered via this design is then used either to rescale the outcome as a change score (Allison 1990) or as a regression control variable (Lin 2013) in the estimation of treatment effects. In what follows, we demonstrate the differencing approach since it is more analogous to the true $SE(\widehat{ATE}_{\text{Standard}})$ introduced above. Change scores are unbiased estimators for a pre-post design, but covariate adjustment may yield more precise estimates (Lin 2013). Appendix D discusses the covariate adjustment approach.

All assumptions for the standard design and ATE remain the same as in subsection 3.1, but now we observe a pre-treatment measure of the outcome for each unit ($Y_{i,t=1}$) in addition to the post-treatment observed outcome ($Y_{i,t=2}$). We make an additional assumption that because $Y_{i,t=1}$ is measured before treatment, its value does not depend on the potential outcomes: $E[Y_{i,t=1}] = E[Y_{i,t=1}|Z_i = 1] = E[Y_{i,t=1}|Z_i = 0]$.

The estimator for the ATE is analogous to $\widehat{ATE}$, but now replacing the outcome of interest with the difference in outcomes before and after treatment:

$$\widehat{ATE}_{\text{Diff}} = E[(Y_{i,t=2}(1) - Y_{i,t=1})|Z_i = 1] - E[(Y_{i,t=2}(0) - Y_{i,t=1})|Z_i = 0]. \tag{5}$$

This is the difference in differences estimator. It is also an unbiased estimator of the ATE. In

[8] Appendix B uses a toy data set to show an example of balancing precision and retention when blocking.

the hypothetical case of $Y_{i,t=1}$ being equal to zero across all units, then $\widehat{ATE}_{\text{Diff}}$ is equivalent to $\widehat{ATE}$.

The standard error is (A. S. Gerber and Green 2012, 98):

$$SE(\widehat{ATE}_{\text{Diff}}) = \sqrt{\frac{\text{Var}(Y_{i,t=2}(0) - Y_{i,t=1}) + \text{Var}(Y_{i,t=2}(1) - Y_{i,t=1}) + 2\text{Cov}(Y_{i,t=2}(0) - Y_{i,t=1}, Y_{i,t=2}(1) - Y_{i,t=1})}{N-1}}.$$

The more predictive $Y_{i,t=1}$ is of $Y_{i,t=2}$, the more the variance of the potential outcomes in the numerator of $SE(\widehat{ATE}_{\text{Diff}})$ shrinks relative to the standard design. However, like with block randomization, these benefits of pre-post designs require sample size is not adversely affected. If the denominator decreases as the numerator decreases, the effects of the pre-post design choice on precision are called into question. In other words, if there is potential sample size loss due to implementing a pre-post design, the researcher now needs to balance the components of $SE(\widehat{ATE}_{\text{Diff}})$ when designing their experiment. Thus, like with block randomized designs, these precision gains can be questioned if sample loss accompanies the design choice.

3.4 Combining alternative designs

Block randomized and pre-post designs are not mutually exclusive strategies, and the lines dividing each strategy are blurry. For example, Bowers (2011) considers a case in which pre-treatment outcomes are used for blocking, while Clifford, Sheagley, and Piston (2021) consider quasi pre-post designs in which a covariate that is highly predictive of the outcome is used in place of pre-treatment outcomes.

This means the choice of alternative strategies in this article reflects not only the decision of whether to invest in measuring covariates before treatment, but also what to do with those variables. Block randomization pertains to how units are partitioned into treatment and control groups, constraining the set of potential randomization schemes to those that we have

good reason to believe have lower $SE(\widehat{ATE})$. This implies block randomization performs better for covariates that are expected to correlate with potential outcomes.

Pre-post designs tackle precision from a different angle. Pre-post designs focus on decreasing noise during estimation of treatment effects, thus after treatments have been administered and outcomes have been measured. In this case, the aim is to choose outcomes (or covariates) that are highly predictive of outcomes.

Since the strategies we discuss reflect different reasons to invest in measuring pre-treatment information, a researcher can use both strategies simultaneously by using pre-treatment information to assign treatment within block and to redefine the outcome (or use covariate adjustment). The next two sections illustrate how to navigate the decision to use block-randomization, pre-post designs, both, or neither.

4 Application: Social Media Comments and Belief in Misinformation

We next assess how to balance precision and retention using data from a published experiment. Anspach and Carlson (2020) conduct an experiment to assess the extent to which social media users believe misinformation in the comments of a news post that contains factually-correct information. The authors find evidence that misinformed comments can cause people to retain the incorrect information, despite the correct information being available. Using the authors' replication archive, we imagine how an applied researcher might navigate block randomized and pre-post design decisions in this context.[9]

Anspach and Carlson's experiment has four arms. The first arm was a baseline condition showing participants a full news article that cited Trump's factually-correct approval rating (36%) from a recent poll at the time the experiment was fielded. The second arm showed

[9]The Harvard Dataverse replication archive is located at Anspach and Carlson (2018).

only the news article preview post as it would appear on a Facebook news feed with the factually-correct approval rating in the preview. The third and fourth arms were identical to the second arm but showed a comment invoking an incorrect approval rating of 49% or 23%, respectively. Thus, the final two arms provide liberal and conservative social commentary that communicates incorrect information alongside available factual information in the preview.

After exposure to the treatment post, participants were asked survey questions measuring trust of the news source, cited poll, and the person posting the commentary. Participants were also asked about Trump's approval rating to gauge belief in misinformation. The authors used a convenience sample from Amazon Mechanical Turk, and 953 respondents were randomly assigned across the four treatment conditions.

The Anspach and Carlson (2020) experiment is a useful case study for two reasons. First, the authors conduct what we call a standard design in Table 1. They use post-treatment measures of the outcome and, we assume, complete randomization. Thus, we can simulate how hypothetical alternative designs compare to the standard design. Second, this was a survey experiment conducted with an online sample using the Qualtrics survey software. Block randomization with at least one covariate is simple to implement in this setting. The authors collected pre-treatment information that could have been used to block randomize, such as partisanship, thus we gauge the benefits of alternative design choices in a setting in which they can be realistically implemented.

Moreover, this experiment provides an interesting case study to evaluate pre-post designs because it is constrained from implementing a true pre-post design. The study's outcomes only make sense in the context of the experimental stimuli. The authors could not ask how much the respondents trusted a poll before the respondents had been exposed to it. Therefore, the authors did not collect pre-treatment outcomes that we can use to investigate the precision afforded by a pre-post design. However, this gives us the opportunity to simulate a quasi-pre-post design, where a question similar but not identical to the outcome is asked

pre-treatment (Clifford, Sheagley, and Piston 2021). In this case, the researchers might have asked if respondents trusted election polling in general to use as the pre-treatment measure of their outcome of interest (trust in the poll in the stimuli). In addition to assessing block randomization given the pre-treatment information the researchers collected, we simulate a quasi-pre-post design with a pre-treatment measure of the outcome that is strongly and weakly correlated with the observed outcome.

Taken together, this replication data provides an example of a common decision researchers face—would implementing a pre-post or block randomized design instead have led to a loss in sample size? If so, would any gains in precision from these design choices outweigh losses in precision from attenuation of the sample size?

4.1 Balancing precision and retention

To demonstrate these competing components of precision, we consult Anspach and Carlson's data with two design choices in mind. First, one might expect that partisans would have different responses to learning about Trump's approval rating. Therefore, a researcher deciding how to design this experiment might consider whether block randomizing on one covariate—a pre-treatment measure of partisanship—might be a worthwhile effort to control for this source of variation in potential outcomes and increase precision in $\widehat{ATE}$. Second, a researcher might consider whether asking about trust in polling would be a worthwhile addition to the pre-treatment survey. By asking this survey item, the researcher could implement a quasi pre-post design and expect to increase precision in $\widehat{ATE}$.

To assess the impacts of alternative design choices in the presence of sample loss, we take the authors' reported treatment effects as truth, simulate potential outcomes from their esimated model, and assess how different hypothetical designs would perform in terms of precision. Specifically, the authors model trust in the poll referenced in the news article or preview controlling for several pre-treatment covariates. We use these reported coefficients when

simulating potential outcomes. In other words, we *assume model four in their reported results is the true state of the world*, and simulate potential outcomes according to it:

$$\begin{aligned} Y_{i,t2} = 3.28 - .18 * \text{Preview}_i - .57 * \text{Liberal Comment}_i - .62 * \text{Conservative Comment}_i \\ +.02 * \text{Need for Cognition}_i - .01 * \text{Need for Affect}_i - .05 * \text{Knowledge}_i + \\ .01 * \text{Age}_i + .01 * \text{White}_i + .02 * \text{Education}_i + .01 * \text{Income}_i - .45 * \text{Party}_i + u_i, \end{aligned} \tag{6}$$

where $u_i \sim N(0, 1)$ is an individual-level random error term and the omitted category is the full article condition. We use the authors' data for the following simulation exercise, adding only one simulated pre-treatment covariate. We simulate a pre-treatment measure of the outcome ($Y_{i,t1}$) and vary the amount it correlates with the outcome, $\rho = [.25, .75]$.

For the following simulation, we focus on the average treatment effect of the preview only condition compared to the full article condition. We focus on this $\widehat{ATE}$ because the authors find it does not reach conventional levels of statistical significance, although it is close ($p = 0.064$). With potential outcomes simulated according to Equation 6, we then simulate the different design decisions outlined in Table 1. In all, we assess the following five designs:

1. **Complete + Post-only (Standard design):** Complete randomization and no pre-treatment information used to when estimating $\widehat{ATE}$
2. **Complete + Pre-post:** Complete randomization including the pre-treatment measure of the outcome as a predictor when estimating $\widehat{ATE}$
3. **Block on Partisanship (PID) + Post-only:** Block randomization, blocking only on a three-item indicator of partisanship
4. **Block on outcome + Pre-post:** Blocking on the pre-treatment measure of the outcome and also using it as a predictor when estimating $\widehat{ATE}$
5. **Block on everything + Prepost:** Blocking on all of the covariates used in Equation 6, including the simulated pre-treatment measure of the outcome, and using a pre-post

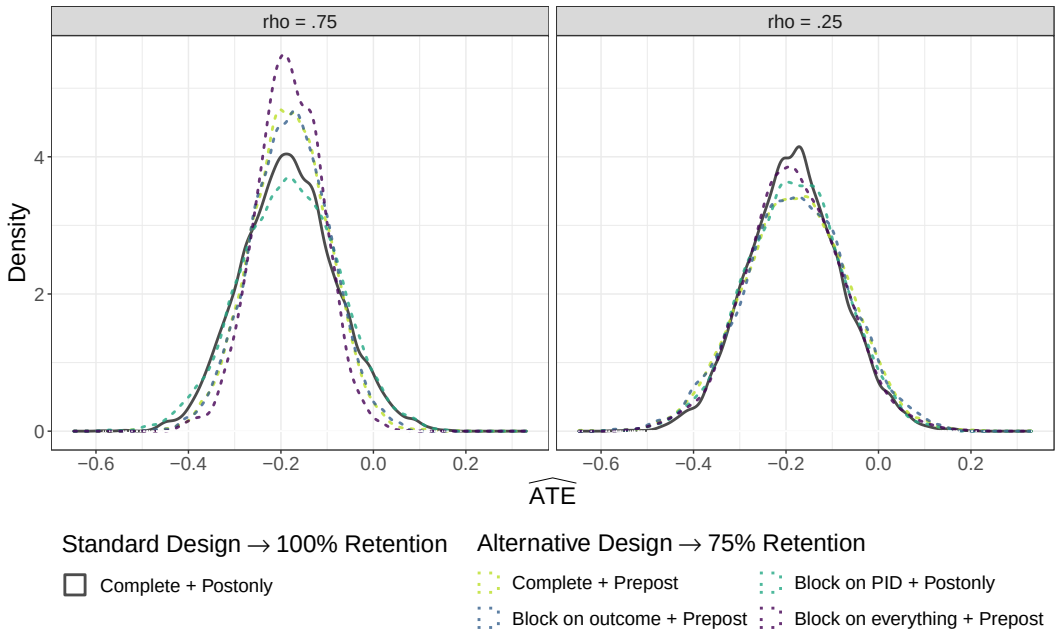

Figure 1: Alternative designs improve precision even with sample attenuation

design

The first is the standard design, and as such, we assume no sample loss. The remaining four designs implement alternatives to the standard design, and we penalize the sample size by assuming only 75% retention as a consequence of implementing an alternative design. The 25% sample loss could be explicit sample loss if the researchers, hypothetically, added so many covariates that the participants dropped from the study from fatigue. Or, we could consider the 25% sample loss as implicit sample loss, if the longer surveys meant they needed to compensate participants more, and their fixed budget required a smaller sample size.

We simulated each of the five designs 1,000 times. Finally, we repeated this procedure twice. Once with a highly predictive pre-treatment measure of the outcome ($\rho = .75$) and once with a weakly predictive measure ($\rho = .25$).

The left plot in Figure 1 displays density plots of $\widehat{ATE}$ for each of the four designs with $\rho = .75$. First, we see that three of the alternative designs do significantly better than the standard design, even though they retain only 75% of responses. Only blocking on

Table 2: Power of standard design and alternatives with sample loss

	N	Rho=.75	Rho=.25
Complete + Postonly (Standard)	100%	0.45	0.45
Complete + Prepost	75%	0.58	0.36
Block on PID + Postonly	75%	0.40	0.40
Block on outcome + Prepost	75%	0.57	0.36
Block on everything + Prepost	75%	0.70	0.42

partisanship does slightly worse that complete randomization. Blocking on just this one covariate, even with a penalty of losing 25% of responses, maintains a similar level of precision as complete randomization with a post-treatment only measure of the outcome.

Moving to the right plot in Figure 1, we can see how these results are contingent upon the pre-treatment information being highly predictive of the outcome. In this plot, the simulated pre-treatment measure of the outcome is only weakly correlated ($\rho = .25$) with the post-treatment measure. Now, the precision gains from incorporating this information cannot overcome the precision losses from losing sample size.

To consider the information in Figure 1 differently, consider the power of the standard design and its alternatives reported in Table 2. The original experiment was underpowered to detect this effect size. However, power increases when blocking on all the pre-treatment information available, even after losing 25% of the sample to do so. This alternative design increases power to .70 relative to .45 under the standard design with a full sample.

Finally, we extend the simulation in Figure 1 to consider *all* combinations of potential sample loss and predictive quality of pre-treatment covariates. For each of the four alternative designs we consider, we simulate $SE(\widehat{ATE})$ across the nearly full range of sample retention (N $= [.05N, .95N]$) and the full range of how well the pre-treatment measure of the outcome predicts the outcome ($\rho = [0, .95]$). Figure 2 visualizes results as a heatmap comparing how alternative designs compare to $SE(\widehat{ATE}_{\text{Standard}})$, where we consider the standard design to be one with 100% sample retention and the pre-treatment and post-treatment measure of the

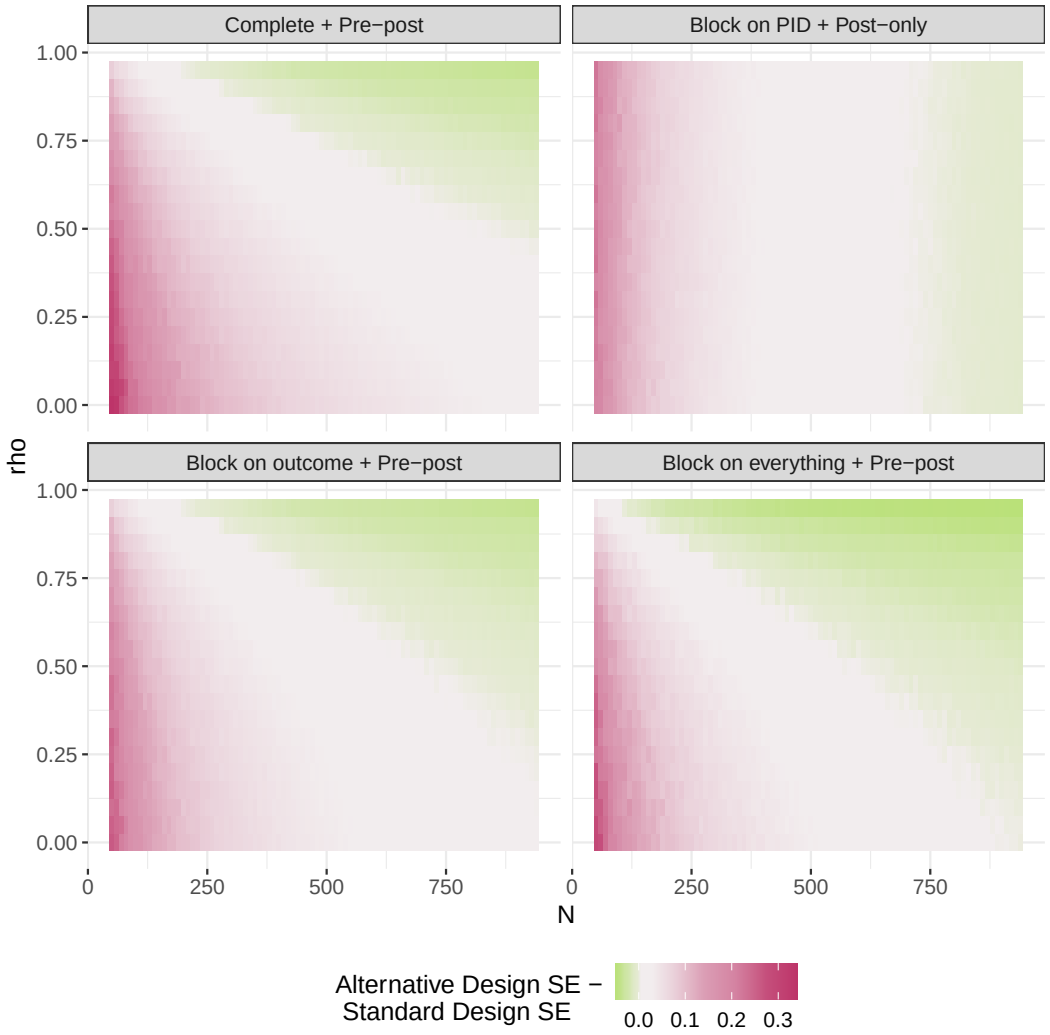

Figure 2: Regions of N and ρ where alternative designs can improve precision

outcome correlated at $\rho = .75$. Green regions represents combinations of N and ρ where the alternative design does better than the standard design, white regions represent where the alternative design does no better or worse, and red regions represent the alternative design does worse than the standard design.

In the bottom left plot, we see blocking on all pre-treatment information available and using a pre-treatment and post-treatment outcome measure correlated at $\rho = .5$, we find that Anspach and Carlson could have precision *gains* even if their sample size dropped to only 680 participants (29% sample loss). In general, in line with the literature's advice, we find that the

more pre-treatment information a researcher utilizes in their design, the more precision gains they can see (bottom right plot). If the pre-treatment information is not highly predictive of the outcome, the gains in precision it brings are not likely to be worth the costs in precision from sample attenuation. We see this in the top right plot where only some sample loss can occur before the decision to block on partisan identity actually harms precision. Finally, we show that if the pre-treatment and post-treatment measure of the outcome are highly correlated, incorporating this information is likely to improve precision, even if it costs the overall sample size.

5 Simulation

5.1 Setting

The previous section uses a published study to illustrate how researchers can balance the competing components of precision in their experimental designs. This section uses simulation to show how researchers can navigate this at the planning stage in a more general setting. The task is to compare the precision in terms of statistical power across the research designs in Table 1.

We simulate two scenarios. First, we conduct an experiment on a sample of $N = 1,000$. We consider one pre-treatment covariate $X_i \sim N(0, 1)$ that is only observed when using block randomization, in which case we construct two blocks depending on whether X_i is positive or negative. This translates to two blocks of similar size. We also consider a pre-treatment outcome $Y_{i,t1} \sim N(0, 1)$ that is only observed in the event of a pre-post design.

We assign a binary treatment to half of the sample via complete randomization. For the designs that include block randomization, treatment assignment is completely randomized within blocks with the same proportion of treated units in each block. The potential outcome under control $Y_{i,t2}(0)$ a standard normal distribution and correlates with $Y_{i,t1}$ with $\rho = 0.8$.

The potential outcome under treatment is $Y_{i,t1}(1) = Y_{i,t2}(0) + \tau Z_i + X_i$ where $\tau = 0.2$ is the true ATE and $Z_i = \{0, 1\}$ denotes treatment assignment. We choose the value of N and τ so that the standard design has middling power, meaning there is room to improve by considering alternative designs.

Since potential outcomes correlate both with the pre-treatment outcome and covariate, we expect any combination of pre-post measurement and block randomization to improve in terms of power in the absence of sample loss. To illustrate this balancing act, we simulate different experiments with varying sample loss rate ranging from 0 to 0.8. We assume two things about sample loss. First, we assume that sample loss happens at random. In some contexts, as in the case of experimental attrition, sample loss may correlate with potential outcomes, which can induce bias to ATE estimation. Appendix E discusses this issue in more detail. In short, we show that this form of correlated sample loss further complicates the balance of precision and retention by adding bias to the mix, yet we argue that one should not worry about this issue any more than what one should worry about experimental attrition in general.

Second, we assume the standard design never suffers from sample loss. Moreover, sample loss is the same regardless of the alternative research design under consideration. In some contexts, this assumption may not be realistic, since the cost of measuring covariates and outcomes may differ. However, this assumption is sufficient to convey sample loss as a function of the proportion of observations the researcher expects to lose.

Our second scenario follows the same setting, but keeps the sample loss fixed at 0.25 while varying N from 100 to $5,000$ observations. This illustrates the case when entertaining an alternative design implies changing the target sample size. For example, adding a baseline survey may force the researcher to study 500 units instead of $1,000$, while still losing a quarter of the sample between waves, leading to an effective sample of $500 - 125 = 375$.

For each combination of parameters, we simulate $1,000$ experiments and estimate the ATE

using the corresponding estimator for each design: The difference in means for the standard design, the difference-in-differences for the pre-post design, the block-size weighted average of the difference in means for the block randomized design, and the weighted average of the difference in differences for the block randomized pre-post design. For each combination of parameters and estimators, we compute statistical power as the proportion p-values smaller than the conventional statistical significance cutoff $\alpha = 0.05$.

5.2 Results

Figure 3 shows the distribution of statistical power for our two simulation scenarios. The left panel shows the statistical power of the standard and alternative experimental designs. Since the standard design (complete randomization + post only) does not suffer from sample loss, its power is constant around 0.6 along the horizontal axis. This serves as a benchmark to evaluate whether it is worth investing in alternative research designs. Holding everything else constant, implementing either a block randomization post-only (block + post only) design or a pre-post design under complete randomization (complete + pre-post) improves upon the standard design as long as the researcher expects to lose less than 30% of the sample. Combining both pre-post outcome measurement and block randomization (block + pre-post) improves precision even further, and its power exceeds the standard design as long as the researcher expects to lose less that 60% of the sample.

The translation of these findings to concrete practices depends on the nature of the application. For example, if investing in alternative designs implies collecting a sample that is half as small as the sample one would collect under the standard design, as in the case of an experiment that requires baseline and endline surveys, then only the block + pre-post design leads to an improvement in terms of precision. Another way to interpret the results is to compare power horizontally. For example, implementing only one of block randomization or pre-post measurement without any sample loss has roughly the same power as an experiment that loses 40% of the sample by using both. Yet another way to interpret the results would be to

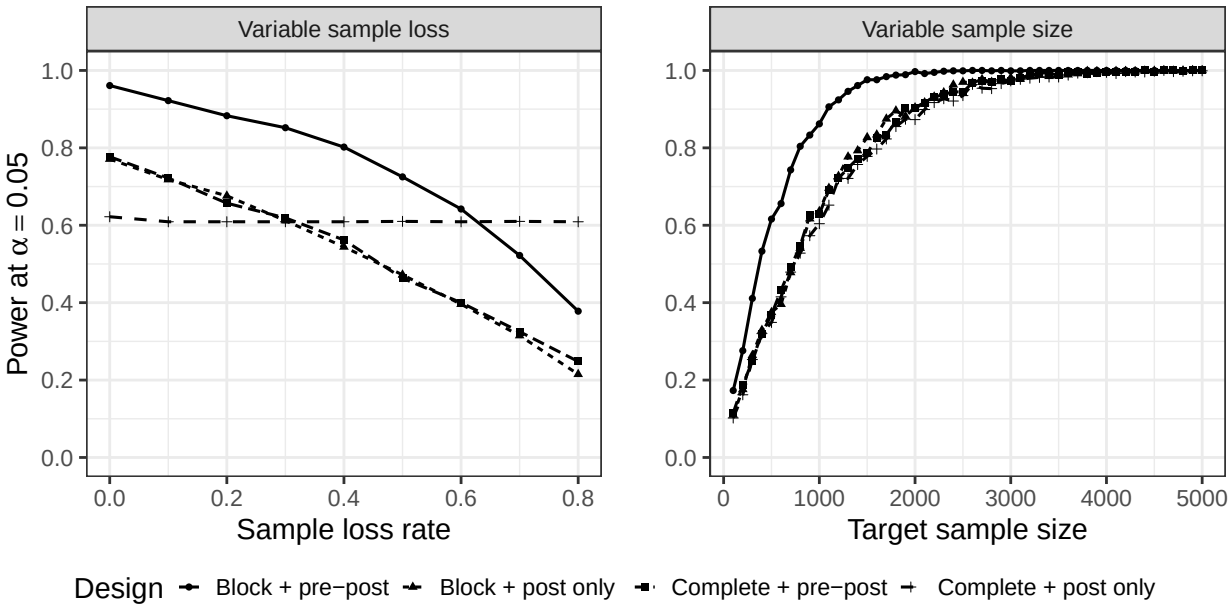

Figure 3: Statistical power for simulated experiments along sample loss rate and target sample size
Note: Each point along the horizontal axis is based on 1,000 simulated experiments.

interpret sample loss in terms of how much smaller of a sample a researcher can afford by investing in a new design. For example, one can afford to collect 60% fewer observations by implementing a block + pre-post design and still achieve comparable levels of precision.

The right panel of Figure 3 shows power for the same alternative designs, but fixing sample loss at 25% and varying the target sample size, which means that the effective sample size is smaller for the alternative designs. For example, a complete + pre-post design with a target sample size of $1,000$ ends up collecting information for $1,000 - 250 = 750$ observations. At this rate of sample loss, the increase in precision from implementing either a complete + pre-post or block + post only is offset by the reduction in the effective sample size. Given these parameters, only the block + pre-post combination leads to a net increase in power. If we focus on the conventional target of 80% power, an experiment that loses 25% of the observations by deviating from the standard design can achieve this with around 800 observations under a block + pre-post design, and with around 1500 observations otherwise, including the standard design.

These findings depend exclusively on the parameters we chose for our stylized simulations, but they convey three types of conclusions that researchers can draw by entertaining the choice of alternative designs at the pre-analysis stage. First, if the goal is to entertain explicit sample loss emerging from the marginal cost of measuring one additional variable, then the application is more likely to exist in the domain of the vertical comparisons in the left panel, and the question is whether one would be willing to sacrifice a small to moderate decrease in sample size to increase precision.

Second, if the goal is to address implicit sample loss from being forced to collect a smaller sample, then the horizontal comparisons in the left panel are more relevant. The question is then how much of the sample loss associated with conducting an additional wave in data collection is tolerable in terms of preserving the target statistical power.

Finally, if the goal of entertaining alternative designs is to minimize data collection costs while preserving statistical power while accounting for both kinds of sample loss, then the comparisons in the right panel provide guideline to determine what would be the minimal target sample to collect under alternative research designs.

6 Conclusion

Previous work proposes deviations from the standard experimental design to improve statistical precision under the assumption that sample size is not affected. This article develops standards to choose among alternative designs under explicit or implicit sample loss. In doing so, we join the conversation on the benefits of simulating experimental designs during the design stage (Blair et al. 2019). Our systematic treatment of the common, competing components of precision highlights how researchers may simulate their experimental design to specifically look for and seek to optimize precision. We hope researchers simulate their designs to understand the extent to which the precision gains of incorporating pre-treatment information into their design in the form of block randomization and/or pre-post measurement withstands any

possible sample size attenuation, perhaps even finding that some sample loss is worth it for large precision gains that can come from these design choices.

This article advances three important conversations in the political science research design literature. First, this article sheds light on how to balance theoretically advantageous design decisions when practical concerns arise. We think it is critical that research unpack and speak directly to best practices, straddling between a statistical understanding afforded by textbooks and a practical understanding of what it takes to implement an experiment. The latter knowledge is acquired through trial and error and conversations with advisors and colleagues, and our article aims to incorporate practical concerns into the public, published conversation on experimental design. Critically, we systematically investigate the competing components of precision rather than rely on anecdotal experience from prior studies. We hope this article encourages more research in this vein.

Second, we shed light on one practical concern that we suspect underlies researchers' hesitancy to implement block randomized, pre-post designs, and other similar innovations in experimental design. Researchers will avoid design alternatives that might prompt *any* explicit or implicit sample loss, fearing the negative consequences on precision and power. In line with this caution, our article shows that blindly implementing theoretically beneficial design choices can have inadvertent consequences when practical concerns are considered. However, researchers' caution may be leaving large precision gains on the table. Following intuition alone, which may steer a researcher toward preserving sample size above all else, is not a good strategy, as we show that non-negligible sample loss resulting from alternative designs can result in large precision gains.

Third, we join an important trend in political science emphasizing the pre-analysis stage of experimentation. Our guidelines do not replace a case-by-case understanding of a design's precision. To facilitate the translation between our simulations and practical application, Appendix A presents a flowchart with steps to consider how alternative designs balance

precision and retention. We hope our findings and guidance lay a path for researchers to understand and consider the competing components of precision in their experiment.

References

Adida, Claire L, Christina Cottiero, Leonardo Falabella, Isabel Gotti, ShahBano Ijaz, Gregoire Phillips, and Michael F Seese. 2022. "Taking the Cloth: Social Norms and Elite Cues Increase Support for Masks Among White Evangelical Americans." *Journal of Experimental Political Science*, 1–10.

Allison, Paul D. 1990. "Change Scores as Dependent Variables in Regression Analysis." *Sociological Methodology* 20: 93. https://doi.org/10.2307/271083.

Anspach, Nicolas M., and Taylor N. Carlson. 2018. "Replication Data for: What to Believe? Social Media Commentary and Belief in Misinformation." Harvard Dataverse. https://doi.org/10.7910/DVN/LQQ5FE.

———. 2020. "What to Believe? Social Media Commentary and Belief in Misinformation." *Political Behavior* 42 (3): 697–718.

Arel-Bundock, Vincent, Ryan C Briggs, Hristos Doucouliagos, Marco Mendoza Aviña, and T. D. Stanley. 2022. "Quantitative Political Science Research Is Greatly Underpowered," July. https://doi.org/10.31219/osf.io/7vy2f.

Berinsky, Adam J, Michele F Margolis, and Michael W Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58 (3): 739–53.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113 (3): 838–59.

Bowers, Jake. 2011. "Making Effects Manifest in Randomized Experiments." In *Cambridge Handbook of Experimental Political Science*, edited by James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia, 459–80. Cambridge University Press. https://doi.org/10.1017/cbo9780511921452.032.

Bowers, Jake, Gustavo Diaz, and Christopher Grady. 2022. "When Should We Use Biased Estimators of the Average Treatment Effect?"

Bowers, Jake, and Thomas Leavitt. 2020. "Causality and Design-Based Inference." In *The SAGE Handbook of Research Methods in Political Science and International Relations*, 769–804. SAGE Publications Ltd. https://doi.org/10.4135/9781526486387.n44.

Box, George EP, William H Hunter, Stuart Hunter, et al. 1978. *Statistics for Experimenters*. John Wiley; sons New York.

Broockman, David E., Joshua L. Kalla, and Jasjeet S. Sekhon. 2017. "The Design of Field Experiments with Survey Outcomes: A Framework for Selecting More Efficient, Robust, and Ethical Designs." *Political Analysis* 25 (4): 435–64. https://doi.org/10.1017/pan.2017.27.

Brutger, Ryan, Joshua D Kertzer, Jonathan Renshon, Dustin Tingley, and Chagai M Weiss. 2020. "Abstraction and Detail in Experimental Design." *American Journal of Political Science*.

Burge, Camille D, Julian J Wamble, and Rachel R Cuomo. 2020. "A Certain Type of Descriptive Representative? Understanding How the Skin Tone and Gender of Candidates Influences Black Politics." *The Journal of Politics* 82 (4): 1596–1601.

Clifford, Scott, Geoffrey Sheagley, and Spencer Piston. 2021. "Increasing Precision Without Altering Treatment Effects: Repeated Measures Designs in Survey Experiments." *American Political Science Review* 115 (3): 1048–65. https://doi.org/10.1017/s0003055421000241.

Coppock, Alexander, Alan S. Gerber, Donald P. Green, and Holger L. Kern. 2017. "Combining Double Sampling and Bounds to Address Nonignorable Missing Outcomes in Randomized Experiments." *Political Analysis* 25 (2): 188–206. https://doi.org/10.1017/pan.2016.6.

Druckman, James N., and Donald P. Green. 2021. "A New Era of Experimental Political Science." In *Advances in Experimental Political Science*, 1–16. Cambridge University Press. https://doi.org/10.1017/9781108777919.002.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and*

Interpretation. WW Norton & Co. https://www.ebook.de/de/product/16781243/alan_s_gerber_donald_p_green_field_experiments_design_analysis_and_interpretation.html.

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, Kevin Arceneaux, Cheryl Boudreau, Conor Dowling, Sunshine Hillygus, Thomas Palfrey, Daniel R. Biggers, and David J. Hendry. 2014. "Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee." *Journal of Experimental Political Science* 1 (1): 81–98. https://doi.org/10.1017/xps.2014.11.

Guess, Andrew M, and Kevin Munger. 2020. "Digital Literacy and Online Political Behavior." *Political Science Research and Methods*, 1–19.

Higgins, Michael J., Fredrik Sävje, and Jasjeet S. Sekhon. 2016. "Improving Massive Experiments with Threshold Blocking." *Proceedings of the National Academy of Sciences* 113 (27): 7369–76. https://doi.org/10.1073/pnas.1510504113.

Imai, Kosuke. 2008. "Variance Identification and Efficiency Analysis in Randomized Experiments Under the Matched-Pair Design." *Statistics in Medicine* 27 (24): 4857–73. https://doi.org/10.1002/sim.3337.

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. https://doi.org/10.1111/j.1467-985x.2007.00527.x.

Kane, John V., Yamil R. Velez, and Jason Barabas. 2023. "Analyze the Attentive and Bypass Bias: Mock Vignette Checks in Survey Experiments." *Political Science Research and Methods*, February, 1–18. https://doi.org/10.1017/psrm.2023.3.

King, Gary, Emmanuela Gakidou, Nirmala Ravishankar, Ryan T Moore, Jason Lakin, Manett

Vargas, Martha Maria Tellez-Rojo, Juan Eugenio Hernandez Avila, Mauricio Hernandez Avila, and Hector Hernandez Llamas. 2007. "A 'Politically Robust' Experimental Design for Public Policy Evaluation, with Application to the Mexican Universal Health Insurance Program." *Journal of Policy Analysis and Management* 26 (3): 479–506.

Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *The Annals of Applied Statistics* 7 (1). https://doi.org/10.1214/12-aoas583.

Lo, Adeline, Jonathan Renshon, and Lotem Bassan-Nygate. 2023. "A Practical Guide to Dealing with Attrition in Political Science Experiments." *Journal of Experimental Political Science*.

Moore, Ryan T. 2012. "Multivariate Continuous Blocking to Improve Political Science Experiments." *Political Analysis* 20 (4): 460–79. https://doi.org/10.1093/pan/mps025.

Moore, Ryan T., and Sally A. Moore. 2013. "Blocking for Sequential Political Experiments." *Political Analysis* 21 (4): 507–23. https://doi.org/10.1093/pan/mpt007.

Nickerson, David W. 2008. "Is Voting Contagious? Evidence from Two Field Experiments." *American Political Science Review* 102 (1): 49–57.

Ofosu, George K, and Daniel N Posner. 2021. "Pre-Analysis Plans: An Early Stocktaking." *Perspectives on Politics*, 1–17.

Pashley, Nicole E., and Luke W. Miratrix. 2021a. "Block What You Can, Except When You Shouldn't." *Journal of Educational and Behavioral Statistics*, July, 107699862110272. https://doi.org/10.3102/10769986211027240.

———. 2021b. "Insights on Variance Estimation for Blocked and Matched Pairs Designs." *Journal of Educational and Behavioral Statistics* 46 (3): 271–96. https://doi.org/10.3102/1076998620946272.