Balancing Precision and Retention in Experimental Design*

Gustavo Diaz† Erin L. Rossiter‡

February 13, 2025

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 when to expect these designs to 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.

*We thank Michelle Dion, Jeff Harden, Geoffrey Sheagley, Natán Skigin, and participants at MPSA 2022 and PolMeth 2022 for helpful feedback. We also thank Amy Brooke Grauley and Shay Hafner for exceptional research assistance. The research involving human subjects was reviewed and approved by the University of Notre Dame Institutional Review Board (Protocol Number: 24-06-8641).

†Assistant Professor of Instruction. Department of Political Science. Northwestern University. E-mail: gustavo.diaz@northwestern.edu

‡Assistant Professor. Department of Political Science. University of Notre Dame. E-mail: erossite@nd.edu

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. One common strategy is to increase sample size if resources permit, even if only in the control group (Gerber, Green, and Larimer 2008). Another main strategy to improve precision is to design the experiment to reduce the variance of outcome measurements. Some designs decisions under this strategy include 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 focus on two designs that rely on *pre-treatment variables* to enhance precision. In particular, we revisit two precision-improving research designs that are promoted as particularly simple and effective: (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 designs may *decrease* sample size, because their implementation requires researchers collect additional information about units before administering experimental treatments. Sample loss from implementing an alternative design may offset promised precision gains in two ways, making it unclear how researchers should navigate these design decisions. First, *explicit sample loss* happens when a study's units 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. Or, participants may drop from the study if a pre-treatment survey is too long or contains too much political content. Second, *implicit sample loss* happens when investing in an alternative design forces the researcher to settle with a smaller sample size, mainly 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 post-treatment. Not all experiments that implement block randomized or pre-post designs may face sample loss, however a design choice's promise of increased precision should be questioned if sample size is adversely affected by it, either explicitly or implicitly.

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. As we will show in Section 4, these designs were implemented in only 15% (32/216) of a sample of

experiments published in the 2022-2023 issues of six political science journals.[1]

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 four perspectives. First, we make the competing components of precision clear, taking a broad view of how sample size can be affected both explicitly and implicitly. Second, we review the current use of these designs in applied experimental work, showing they are infrequently used. We highlight the high stakes involved when deciding whether to implement these designs. The benefits may be large, or the choice may actually harm precision. Third, to help researchers consider the benefits of these designs, we present empirical evidence. We replicate three published experiments, randomizing participants to either a block randomized, a pre-post, or a what we call the "standard" design with complete randomization and post-treatment measurement of the outcomes. Our goal is not to reassess substantive findings or conclusions. Instead, randomizing participants to counterfactual designs provides us leverage on the consequences of implementing these *in practice* for sample loss, increased precision, and more. Fourth, we conduct six simulation studies using data from published experiments. These simulations illustrate how a researcher can entertain alternative designs and potential sample loss before conducting an experiment.

Taking all of our evidence together, 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 evidence and simulations show that highly predictive blocking covariates and pre-treatment outcome measures can produce precision gains that *withstand non-negligible sample loss*. In other

[1]This evidence comports with the findings of Clifford, Sheagley, and Piston (2021) who find only 18% (12/67) of a sample of experiments deviate from measuring only post-treatment outcomes.

words, perhaps running counter to what a researcher would expect, sacrificing sample size in order to implement block randomization or pre-post measurement may greatly increase precision. 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. In the final section of the article, we summarize how our findings can inform future experiments. By developing guidelines to determine whether the investment in alternative designs is worth it, we further expand researchers' ability to implement experimental designs that will detect non-zero treatment effects when they exist across a broad range of applications, and particularly in contexts when a researcher has a limited budget.

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 often motivate researchers to entertain alternative research designs. Researchers can deviate from the standard design by choosing an alternative randomization procedure or an alternative timing of the outcome measurement. Both strategies depend on the availability of pre-treatment variables to be implemented. For example, one would need access to crime statistics to assign treatments independently across low and high crime areas. When these variables are not available, implementing alternative designs comes with the additional cost of collecting them.

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 Randomized Design

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

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

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).

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 collected before treatment assignment and used in estimation of treatment effects. In some contexts, pre-treatment outcomes need to be measured in a separate wave than treatment administration. For example, interventions in field experiments may not involve a survey, so measuring pre-treatment outcomes requires a separate wave to do so. In other contexts, such as many survey experiments, pre-treatment outcomes can be collected, treatment can be administered, and post-treatment outcomes can be measured all in the same wave.

Pre-treatment outcomes can be 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.

2.3 Explicit and Implicit Sample Loss

While these design choices have clear 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. While the literature's recommendation to use these designs assumes sample size is unaffected by the decision to implement them, in practice, researchers often run into contexts where that is unlikely to be the case. This leaves 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.

First, we refer to "explicit sample loss" as circumstances when the sample is already defined and units that 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.[2] 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 for 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

[2]Block randomization may invoke explicit sample loss because matched-pair designs (two treatment conditions and two units per block) requires an even number of units, thus dropping one unit if there were an odd number of units. Moreover, multivariate continuous blocking (Moore 2012) cannot be implemented with missing data, possibly leading a researcher to drop incomplete cases to use key blocking covariates that include missingness.

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.[3] 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.

3 Balancing Precision and Retention Under Alternative Designs

In this section, describe the standard experimental research design and illustrate how implementing an alternative design to increase precision requires a researcher balance sample retention concerns, as well.

[3]This follows from the cost calculator at https://www.prolific.co/old/pricing, assuming the default hourly rate of USD$10.54 per respondent as of February 21, 2023.

3.1 The Standard Experimental Design

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 ($Y_i(0)$). We assume that potential outcomes satisfy SUTVA and excludability, and that treatment is randomly assigned.

The first defining feature of what we will call 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.

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 (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 particularly simple and 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 $Var(Y_i(1))$.[4]

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 $SE(\widehat{ATE})$.

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

3.2 Block Randomized Design

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 $\widehat{ATE}_{\text{Block}}$.

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

Like the $\widehat{ATE}_{\text{Block}}$, the true standard error is a weighted average of within-block standard errors (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.

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 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.

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 C 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: $Y_{i,t=1} = E[Y_{i,t=1}] = E[Y_{i,t=1}|Z_i = 1] = E[Y_{i,t=1}|Z_i = 0]$.

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 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 (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. 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 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).

4 Current Use of Alternative Designs to Increase Precision

How common are alternative designs in political science? Many applied experimentalists may follow statistical advice to use pre-post and/or block-randomized designs to increase precision. Alternatively, these designs might be rare, yet experiments still achieve sufficient precision. We review the applied experimental literature to answer whether important precision gains seem likely from increased use of these designs.

To conduct this review, we hand coded features of experiments in articles published in 2022-2023 in six political science journals. We reviewed *The American Political Science Review*, *The American Journal of Political Science*, and *The Journal of Politics* as the field's major general interest journals. These journals publish experiments that come from all subfields, implement different experimental designs, and reach a wide audience. However, these experiments may not reflect the modal design in the discipline. Most notably, these experiments may have had larger budgets to allow for pilot testing, larger samples, and other design features to increase precision. Therefore, we also review three additional journals. We include *Political Behavior* and *Comparative Political Studies* as two subfield journals where experiments are commonly published, and we include *The Journal of Experimental Political Science* as the discipline's journal devoted to experimental research. This set of journals allows us to survey how applied experimentalists currently balance precision and retention concerns.

We collect all full-length articles with at least one original randomized experiment, resulting in 227 articles and 366 unique experiments. We sampled approximately 50% of these articles for our analysis. Our sample has 121 articles and 216 unique experiments. For sampled articles, we read the methods, results, and discussion sections to hand code *each experiment* separately for several experimental design concepts. We discuss the details of our inclusion criteria and hand coding procedure in Appendix D.

Table 2 shows our results. We found 85% of the experiments did not use pre-post or block randomized designs. Only 5% used a pre-post design, 7% used a block randomized design, and less than 3% used pre-post and block randomization together. These results demonstrate that the alternative designs of interest in this article are infrequently used by experimental political scientists.

Table 2: Current Use of Alternative Designs to Increase Precision

	Prevalence	Median (IQR) Sample Size	Median (IQR) Arms
Neither	85% (184)	1382.5 (801, 2495.5)	4 (2, 6)
Only prepost	5% (10)	5068 (1057.5, 6042)	3 (2, 3)
Only block randomization	7% (16)	1545.5 (128, 2768)	2 (2, 3)
Both prepost and block randomization	3% (6)	1230 (734.5, 2448.5)	3 (2.25, 3.75)

Note: The second column shows the percentage of designs used in a sample of 216 experiments from articles published in 2022-2023, with the number of experiments in parentheses. The third and fourth columns show the median sample size and number of experimental arms per type of design, with the interquartile range in parentheses.

However, 92% of the experiments that did not implement either design mentioned having pre-treatment covariates. We do not know whether these covariates were well-suited for pre-post or block randomized designs. However, having pre-treatment covariates suggests that many designs likely had the *opportunity* to use this information or collect additional information that could be used in a pre-post or block randomized design.

We also cannot quantify how collecting different or additional pre-treatment covariates might lead to implicit or explicit sample loss. However, we can characterize the high stakes of this decision. We collected the sample size and the number of experimental conditions in each experiment. Among the experiments that did not implement pre-post or block randomized designs, the median sample size was 1382 and the median number of experimental conditions was 4. While we cannot say the median experiment is underpowered, it is also not a foregone conclusion that a design with these features would be sufficiently powered, especially given meta-analytic evidence showing political science experiments are greatly underpowered (Arel-Bundock et al. 2022). This characterizes the puzzle facing researchers when considering the implementation of pre-post and block randomize designs. The median political science experiment likely has room to make critical precision gains that will enhance the experiment's ability to detect true treatment effects with greater certainty. At the same time, any design decision that would sacrifice sample size by asking additional questions (incurring implicit sample loss) or by introducing conditions that increase unit's likelihood of dropping from the study (incurring explicit sample loss) ought to be carefully considered. This article guides applied researchers through balancing these competing components of precision.

5 Benefits of Designs to Increase Precision in Applied Settings: Experimental Evidence

To assess whether and to what extent alternative experimental designs increase statistical precision even when incurring sample loss, we conducted a preregistered replication of three published experiments.[5] For each experimental replication, we conducted *our own* experiment. We randomized whether participants were assigned to the standard design (complete randomization and post-treatment outcome measurement only), a pre-post design (with complete randomization), or a pre-post and block randomized design. Our replication experiments provide evidence to address several key questions about implementing alternative designs in practice.

Before presenting this evidence, we describe the replication exercise. We follow the framework proposed by Harden, Sokhey, and Wilson (2019) for selecting included studies. We summarize each step of this framework, including defining a population, constructing a sample representative of the population, and defining the quantities of interest from the replication exercise.

First, we define the population as all published, original randomized experiments in political science where the goal of the experiment is a substantive finding. To sample from this population, we utilize the data we collected for our hand coding exercise described in Section 4. In Section 6, we discuss a simulation exercise that randomly samples from this list of published experiments. In this replication, we hope to generalize to the population while maintaining feasibility of the original data collection, so we take a fine-tuned approach to choosing experiments to replicate.

Table 3 lists the three articles we chose.[6] We replicate Dietrich and Hayes (2023), Bayram and Graham (2022), and Tappin and Hewitt (2023), which we refer to as DH, BG, and TH, respectively. These experiments cover different subfields and are representative of the distribution of experimental conditions and observations from our population. DH and BG were single-wave studies, like most survey experiments. TH allowed us to assess a multi-wave study where the first wave solely collected pre-treatment covariates, like in many field experiments. Finally, it was not possible to implement a pre-treatment measure of the outcome in the DH replication, as the main outcome

[5] The preregistration is available at https://doi.org/10.17605/OSF.IO/XJ35P.

[6] We also preregistered the specific outcome and treatment effect from the original study that would be our focus in this exercise.

was a reaction to the experimental stimuli. Therefore, we used a quasi pre-post measure of the outcome, while the other two replications allowed for a pre-treatment measure of the outcome. Taken together, these experiments vary key features that affect precision, allowing our evidence to have some generalizability to the population.

We also chose this set of experiments because they have different motivations for balancing precision and sample retention. For DH, one of their hypotheses pertains specifically to how African American constituents respond to rhetoric from members of Congress. Addressing precision concerns via increasing sample size is not always an option in this context, as survey providers often have a limited number of Black or African American respondents. Researchers who are interested in small or hard-to-reach populations therefore have a motivation to design their experiments to increase precision through other means. We chose BG as a standard example of how an experiment can always consider alternative designs with an eye toward increasing precision, and thus their confidence in their conclusions, even when power analyses at the design stage suggest the experiment achieves conventional levels of power. Finally, TH assesses whether the effect of party cues on issue attitudes persists over time. We chose to replicate this experiment because treatment effect durability is an important concern in experimental political science—particularly in survey experiments—to contextualize the political relevance of treatment effects (Gaines, Kuklinski, and Quirk 2007). Moreover, to assess durability, researchers must consider the cost of remeasuring outcomes and whether their design will be powered to detect treatment effects immediately *and* in the future when attrition has occurred and the effect has likely decayed.

For each replication study, we assess how alternative designs affected sample loss and precision relative to the standard design. Table 4 outlines how we implemented the designs and the number of participants randomized to each design per study. DH and BG were both single-wave survey experiments. DH and BG asked 7-8 survey items in the standard design, and 3 and 4 additional items, respectively, in the alternative designs to collect a pre-treatment measure of the outcome and predictive covariates. The standard and pre-post design assigned participants to treatment conditions using complete randomization, while the block randomized design assigned treatment within 15 and 6 unique blocks, respectively, based on a predictive covariate and the pre-treatment measure of the outcome. Finally, TH was a 3-wave study. Wave 1 asked 6 demographics in the

Table 3: Key Features of Sample Articles Representing Population

	Dietrich and Hayes (2023)	Bayram and Graham (2022)	Tappin and Hewitt (2023)
Subfield	American politics	International relations	American politics
Number of arms	8	5	2
Number of observations	515	1000	775
Number of waves	1	1	3
Type of prepost	Quasi prepost	Prepost	Prepost
Primary precision concern	Hard to reach population	Increased confidence	Detecting effect persistence

standard design, and an additional 12 items in order to implement the alternative designs. Wave 2 administered the treatment using complete randomization in the standard and prepost design. For the block-randomized design, we implemented multivariate continuous blocking between Wave 1 and 2 using demographics, predictive covariates, and pre-treatment measures of the outcomes collected in Wave 1 (Moore 2012; Moore and Schnakenberg 2023). All studies were survey experiments conducted on CloudResearch Connect. More details on the implementation are available in Appendix E and our preregistration.

In sum, while only a few questions were added to implement the alternative designs, doing so significantly increased survey length (43%, 50%, and 110%, respectively). The alternative designs also added political content pre-treatment. This allows us to asses several concerns that may arise when implementing alternative designs in practice, which we turn to next.

Table 4: Implementation of the Standard and Alternative Designs in Each Replication

	Study 1: DH			Study 2: BG			Study 3: TH		
	N	Survey items	Randomization	N	Survey items	Randomization	N	Survey items	Randomization
Design 1: Standard design	426	6 demographics 1 stimulus 1 outcome	Complete	1008	6 demographics 1 stimulus 1-2 outcomes (branched item)	Complete	752	6 demographics (Wave 1) 0-10 stimuli (Wave 2) 5 outcomes (Wave 3)	Complete, start of Wave 2
Design 2: Prepost	427	Standard design +3 items	Complete	1010	Standard design +4 items	Complete	780	Standard design +12 items	Complete, start of Wave 2
Design 3: Prepost & block randomized	429	Standard design +3 items	Block randomization, 15 blocks formed from: •Predictive covariate (3 levels) •Quasi pre-treatment outcome measure (5-pt scale)	1006	Standard design +4 items	Block randomization, 6 blocks formed from: •Predictive covariate (3 levels) •Pre-treatment outcome measure (binary)	767	Standard design +12 items	Multivariate continuous blocking prior to Wave 2 with all Wave 1 covariates, including: •Pre-treatment outcome measures •Demographics •Predictive covariates

5.1 Explicit Sample Loss

First, did alternative designs, requiring longer surveys with more political content, incur greater explicit sample loss? Appendix Table E2 shows detailed results. DH and BG—both single-wave studies—featured little explicit sample loss (<1% per design) that was not statistically distinguishable between the standard and alternative designs.

To complete the TH study, respondents had to return to take three survey waves across 1-1.5 weeks. It follows that TH had more explicit sample loss—13.0%, 14.2%, and 20.7% of units dropped at some point in the standard, pre-post, and pre-post with blocking designs, respectively. The block-randomized design featured more explicit loss because we implemented multivariate continuous blocking, utilizing all pre-treatment covariates. The block randomization procedure does not allow missingness in blocking covariates, so we excluded any observation with missingness in these covariates prior to block randomization and treatment randomization. This decision excluded 7.5% of observations, which explains why this design features more overall sample loss.[7]

In sum, implementing pre-post and/or block randomized designs did not increase the rates of units dropping by their own choice, an encouraging result. However, depending on the researchers' choices when implementing multivariate continuous blocking in a multi-wave setting, this design may incur additional explicit sample loss. This sample loss is pre-treatment, thus the researcher is not risking biased treatment effect estimation.

5.2 Differing Sample Composition

It may be a concern to researchers that implementing an alternative design that requires a lengthy pre-treatment battery may prompt units to drop from the study in a way that alters the *sample* with which the researcher estimates treatment effects. We find little to no evidence of this in our replications. As just discussed, DH and BG had trivially small sample loss and thus no difference in sample composition. However, in the TH replication, we found sizable sample loss throughout the three-wave study in all three designs. We use TH to understand whether implementing alternative designs results in different samples.

[7]Rather than exclude observations with missingness, a researcher could use imputation for missing values or exclude covariates with missingness from their blocking algorithm.

Why might different samples result from alternative designs? First, survey fatigue could cause units to drop in alternative designs in ways that would make the sample different from the standard design, such as if fatigue was related to education or age. Second, asking many *political* items may cause a certain kind of person to drop from the study at a higher rate than if these items were not asked, such as people who have a strong aversion or disinterest in politics. Third, as in the TH replication, if observations are excluded due to missingness in blocking covaraites, and missingness is related to characteristics such as an aversion to politics or inattentiveness, implementing multivariate continuous blocking could affect sample composition.

We investigate whether an observation being included in the sample or not is differentially predicted across designs by any of the pre-treatment covariates we collected (24 demographic indicators).[8] We find no evidence of a difference between the pre-post design's sample and the standard design's sample (Appendix Table E3). We find only one instance where the block randomization with pre-post sample differs from the standard design's sample.

In sum, we find very little evidence that implementing alternative designs that require more pre-treatment items may cause the samples with which we estimate treatment effects to differ in important ways. However, this evidence is limited as we can only assess the pre-treatment covariates we observed, and the number of items was limited for budgetary reasons. Nevertheless, this evidence provides reassurance that sample loss from alternative designs would not result in meaningfully different samples in similar contexts.

5.3 Differential Post-Treatment Attrition

Next, we turn to examining post-treatment attrition. Two forms of post-treatment attrition pose a concern when implementing an alternative design. First, will alternative designs cause more post-treatment attrition? We fail to find evidence that post-treatment attrition rates are different between the alternative designs and the standard design (final column in Appendix Table E2). Second, will alternative designs cause attrition that differs between treatment and control groups? This form of attrition is an important concern as it can compromise the estimation of unbiased

[8] We also assess whether sample inclusion between the two alternative designs is related to the political measures we collected, and we find no evidence that sample composition differs between alternative designs (Appendix Table E4).

treatment effects. To assess this question, we examine if differential attrition across experimental arms is explained by different covariates between the alternative and standard designs. We can only investigate attrition in the TH replication since it was trivially small in the other replications. We find little evidence that pre-treatment characteristics affect attrition across treatment arms differently in the alternative designs relative to the standard design (Appendix Figure E3). Only 4 out of 44 demographic indicators examined had patterns of attrition across treatment arms that differed between the alternative and standard designs. In fact, in 3 of the 4 cases, the covariate is associated with participants attriting *less* from the alternative design's treatment than the standard design's treatment. In sum, we encouragingly find little evidence of differential post-treatment attrition across alternative designs in this replication exercise, again with the caveat that we can only empirically investigate the characteristics we measured.

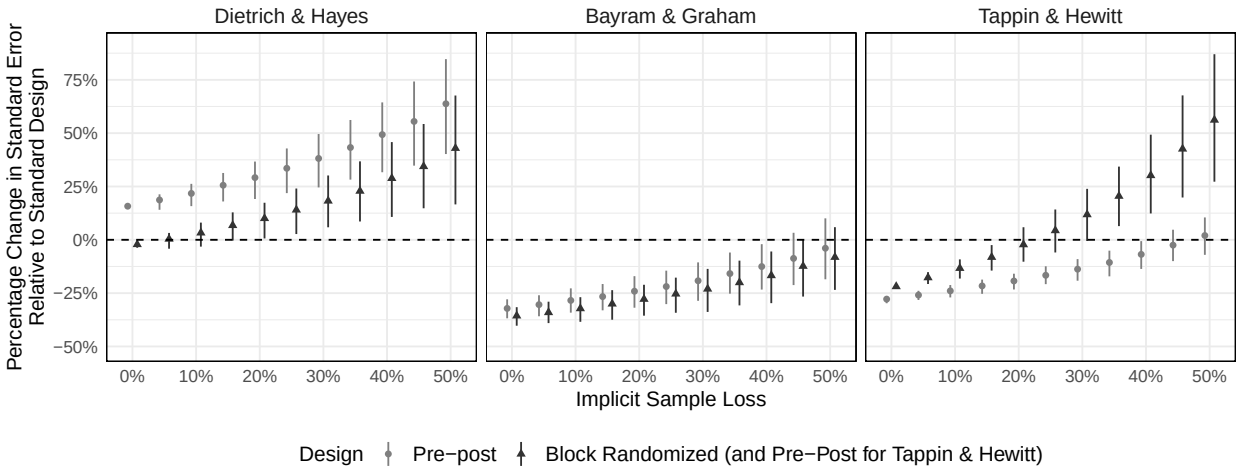
5.4 Implicit Sample Loss

Finally, we use the replications to ask, what are the implications for precision when implicit sample loss occurs as a result of implementing an alternative design? We assess this via simulation.[9] Because of small variation in the number of participants assigned to each design (see Table 4), we begin the simulation with all three designs at the same sample size per study. We then randomly omit observations from the alternative designs to simulate implicit sample loss, re-estimate treatment effects and standard errors, and assess how statistical precision of alternative designs *with* sample loss compares to the standard design *without* sample loss. We vary the amount of implicit sample loss from 0 to 50% in increments of 5%. We conduct 1000 random draws of data per sample size.

Figure 1 visualizes the results. The x-axis shows the hypothetical amount of implicit loss incurred by implementing an alternative design, varying from 0 to 50%. The y-axis shows the percentage change in the estimated standard error of the alternative design *with* sample loss relative to the standard design *without* sample loss. The first panel shows results for DH. Without sample loss, block randomization (dark gray triangles) provides for slightly more precise estimation than the

[9]We use simulation because, within a given study, the cost to implement the three designs was held constant. If a researcher were choosing just one design to implement, a longer survey would cost more than a shorter survey, reducing the sample size on a fixed budget (i.e., implicit sample loss).

Figure 1: Effects of Implicit Sample Loss on Precision



Note: Figure visualize the effects of implicit sample loss on precision. The x-axis shows the amount of implicit loss incurred by implementing an alternative design, varying the loss from from 0 to 50% of the sample. The y-axis shows the percentage change in the estimated standard error of the alternative design with sample loss relative to the standard design without sample loss (with 95% confidence intervals). Results for the prepost and block randomized alternative designs are shown with light gray and dark circles, respectively.

standard design. At 5% sample loss, the precision afforded by block randomization is no different than the precision afforded by the standard design at full-sample. This suggests there is no harm to precision if implementing block-randomization that requires sacrificing 10% of the sample or less to afford it. With 10% sample loss or more, block randomization begins to do worse than the standard design at full sample size.

The results are more surprising for the pre-post alternative design. We find that the pre-post design, even with no implicit sample loss, has an approximately 15% larger standard error than the standard design, and it only grows with sample loss. We believe this occurred for two reasons. First, this replication was the smallest we conducted, with about 430 observations assigned to 8 experimental conditions per design, and the particular treatment effect we preregistered analyzing only has about 210 participants per design. Ironically, the small sample sizes are likely associated with a large sampling distribution of treatment effect estimates, increasing the probability the three designs are not good counterfactuals for each other. Second, we implemented a quasi pre-treatment measure of the outcome for this design, and it was only weakly correlation with the outcome (r=0.28).[10] Taken together, the quasi pre-post design, even with a weak correlation,

[10]The outcome in the DH replication measured a reaction to the experimental stimuli, specifically if they approved of what the legislator said in a speech discussing civil rights that either did or did not employ civil rights symbolism.

should improve precision relative to the standard design. Nonetheless, the pre-post design was *less* precise, likely due to sampling variability.

The second panel shows results for BG. Here, pre-post and block randomization provide large precision gains (over 30%) when no sample loss is incurred. In fact, both alternative designs improve precision of the estimated treatment effect, even with a sample size half as large as the standard design. The pre- and post-treatment outcomes feature strong correlation (r=.70). Asking this item pre-treatment and incorporating it into the design via pre-post estimation or block randomization provided sizable precision gains. The block randomized design used a second pre-treatment covariate to compose blocks; however, using this additional information did not result in precision gains distinguishable from the pre-post design.

The third panel shows results for TH. Again, both pre-post and block randomized designs have sizable increases in precision (25%) relative to the standard design when there is no sample loss. Even at 20% sample loss, block randomization has slightly better precision than the standard design with no loss. Pre-post performs even better in this context. The pre- and post-treatment outcomes feature strong correlation (r=.69). Like with BG, the pre-post design is achieving similar precision with half the sample size as with standard design. In sum, if a researcher did not wish to prime participants by asking a pre-treatment item in the same wave as the post-treatment outcome measurement, or if a researcher must implement pre-treatment outcome measurement in a separate wave for logistical reasons, this replication shows that using pre-treatment information provides precision gains that could withstand sizable sample loss (near 50%).

In summary, this evidence shows that block randomized and pre-post designs provide significant precision gains over the standard design that can withstand large sample losses when pre-treatment information has a strong correlation with the outcome. When this correlation is weak, such as in the DH replication, precision gains that offset sample loss quickly diminish.

Because we could not measure this pre-treatment, we asked a quasi pre-treatment outcome instead, asking "How important is it to you that current members of United States Congress publicly address their commitment to civil rights?" We expected this to be predictive of the outcome by capturing personal importance of such actions.

6 Benefits of Designs to Increase Precision in Applied Settings: Simulation Evidence

To complement our experimental evidence, we next use simulations that allow us to better assess the kinds of pre-treatment and blocking covariates alternative designs ought to use to increase precision that offsets any incurred sample loss.

Rather than simulate entirely fabricated data, our simulations use data from a set of published experiments. We randomly sample six experiments from sample described in Section 4. To increase the generalizability of our findings to political science experiments, we randomly sampled one experiment from each of the six journals included in our review.[11] Table 5 presents key features of the six experiments. As intended, our sample includes different research topics, areas of the world, sample sizes, and number of experimental arms. The ratio of survey and field experiments is also a good reflection of the distribution in our target population.

Each experiment in Table 5 constitutes a separate simulation. Broadly, the first step is to use the original study's data to model the relationship between the outcome, treatment, and covariates. We then simulate potential outcomes from this model, administer the standard and alternative designs, and assess each design's statistical precision.

To generate an assumed true model of the world, we first narrowed our focus to one treatment effect of interest. We did so at the preregistration stage, blind to the data. In experiments with multiple treatment arms, we preregistered a treatment effect for each study that was central for the original article's argument. Similarly, if multiple outcome variables were used, we chose to focus on one of the main outcomes of interest.

After posting our preregistration, we used the replication data to model the outcome as a function of (1) actual treatment assignment in the experiment, (2) a simulated pre-treatment measure of the outcome, (3) a simulated pre-treatment blocking covariate, and (4) a set of actual pre-treatment covariates collected in the original study. For the simulated variables, we vary the extent to which they correlate with the outcome ($r \in [0.25, 0.50, 0.75]$). For the actual set of pre-treatment covari-

[11] See the preregistration materials at https://doi.org/10.17605/OSF.IO/KPWY6 for specific details on our sampling strategy and inclusion criteria.

Table 5: Experiments Sampled for Simulation

		Original experiment				Simulation		
Study	Article	Type	Arms	N	Country	n	Blocking Covariates	Predictiveness
1	Galasso et al (2023)	Survey	6	2971	Italy	946	7	Low
2	Manekin and Mitts (2022)	Survey	12	3013	United States	2784	3	High
3	Lyon (2023)	Survey	3	1029	Uganda	561	4	Low
4	Goerger et al (2023)	Field	2	2942	United States	2712	5	Moderate
5	Curiel et al (2023)	Field	2	275	Colombia	275	7	Low
6	Simas (2022)	Survey	2	1176	United States	1175	2	Moderate

Note:
We assess the predictiveness of blocking by estimating OLS regressions for each outcome against the corresponding covariates. We consider predictiveness as "high" if all covariates have large coefficients, "moderate" if only some variable have large coefficients, and "low" if all coefficients are small. See Appendix Table F8 for results.

Table 6: Designs Included in Simulation

Design	Description
1. Complete + post-only (Standard design)	Complete randomization; pre-treatment outcome is not measured or used when estimating $\widehat{ATE}$
2. Complete + pre-post	Complete randomization; pre-treatment outcome is measured and used as a predictor when estimating $\widehat{ATE}$
3. Block on one covariate + post-only	Block randomization using one pre-treatment covariate; pre-treatment outcome is not measured or used when estimating $\widehat{ATE}$
4. Block on one covariate + pre-post	Block randomization using one pre-treatment covariate; pre-treatment outcome is measured and used as a predictor when estimating $\widehat{ATE}$
5. Block on all covariates + post-only	Block randomization using all selected and simulated pre-treatment covariates; pre-treatment outcome is not measured or used when estimating $\widehat{ATE}$
6. Block on all covariates + pre-post	Block randomization using all selected and simulated pre-treatment covariates; pre-treatment outcome is measured and used as a predictor when estimating $\widehat{ATE}$

ates collected in the original study, we consulted the article and replication archive documentation to preregister a set of covariates we expected to be best predictive of the outcome, but without conducting any analyses, to mimic a researcher at the design stage as much as possible.[12] Next, we assumed true model as the data generation process for potential outcomes. Finally, we then simulate treatment assignment and treatment effect estimation under six different research designs,

[12]Appendix F explains deviations from our preregistration and reports the results as originally preregistered. We found that the majority of variables we designated blocking and quasi outcome covariates had very low correlation with the outcome (Appendix Table F8). This is understandable as they were not originally included in the designs for these purposes. This meant our preregistered results did not reflect the context where a researcher is choosing the best variables for alternative designs. Therefore we instead simulate varying degrees of predictiveness of pre-treatment information to better capture the relationships researchers are likely to be working with. The lack of predictive pre-treatment information in our sample highlights that researchers are not necessarily currently collecting covariates that are good candidates as pre-treatment measures for pre-post designs, and must consider adding measurement of pre-treatment outcomes to their studies.

shown in Table 6.

Our simulations assume that the standard design does not incur sample loss. For every other design, we varying the degree of sample loss by randomly dropping a proportion of units between 0 and 0.5. We simulate 1,000 hypothetical experiments for every design, degree of correlation between the pre-treatment outcome or blocking covariate, and degree of sample loss.

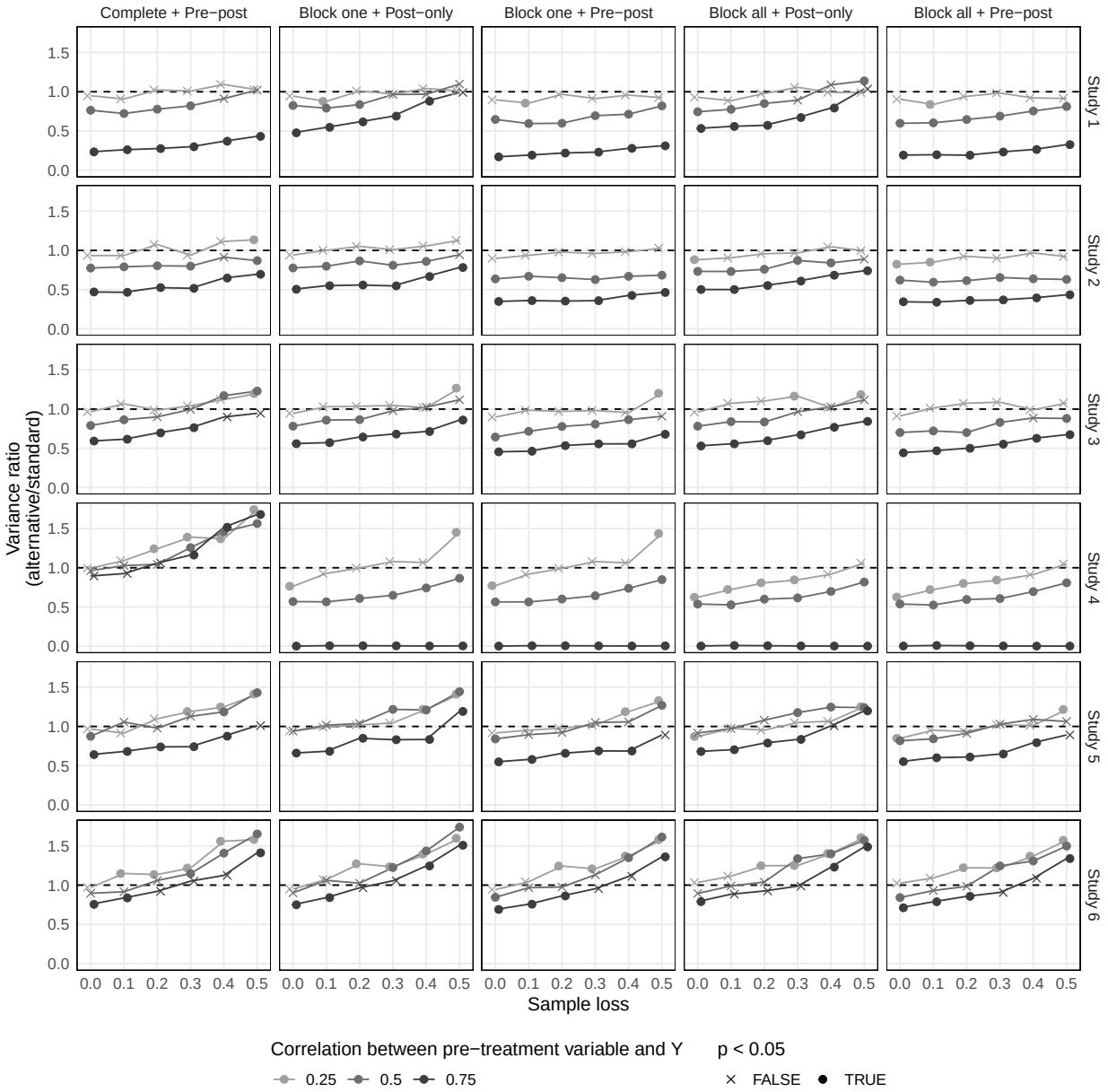
Figure 2 presents the results. We conduct F-tests to compare the alternative design in each column against the standard design. The vertical axis presents the test statistic as the ratio of variance of the simulated $\widehat{ATE}$'s between a given alternative design under sample loss and the analogous standard design without sample loss. If this statistic is distinguishable from one, the evidence suggests the two variances are different from each other. Because the standard design does not incur sample loss, its variance remains unchanged along the horizontal axis per plot. Therefore, test statistic values smaller than 1.0 indicate that the variance of the simulated $\widehat{ATE}$'s is smaller under the alternative design, which suggests that the alternative design is preferable even after incurring sample loss. For example, a value of 0.5 suggests the variance of the alternative design is half the variance of standard design. Conversely, test statistic values larger than 1.0 suggest that the degree of sample loss undermines any precision gains from implementing an alternative design.

In all six studies, including a pre-treatment outcome or covariate that weakly correlates with the outcome ($r = 0.25$, lightest gray line) is not enough to overcome the decrease in precision from sample loss. We only see a benefit in alternative designs with weak pre-treatment information once we incorporate additional blocking covariates in study 2, but these precision gains only withstand up to 10% sample loss before the variance in the alternative design becomes indistinguishable from the standard design.

Moderately correlated variables ($r = 0.50$) are sufficient by themselves to improve precision in studies 1 through 4. In this group, the worst performance lies in study 3, where precision gains only withstand up to 20% sample loss under the Complete + Pre-post design. The best performance lies in studies 1, 2, and 4, where all feature precision gains over the standard design even with up to 50% sample loss under the Block one + Pre-post design column.

As expected, performance only improves when we include highly correlated variables ($r = 0.75$).

Figure 2: Precision Gains from Alternative Designs Under Varying Predictiveness of Pre-treatment Information and Degrees of Sample Loss



Note: Columns represent alternative designs, rows indicate studies in the order listed in Table 5. Test statistics are calculated based on 1,000 simulated experiments for each study, research design, degree of sample loss, and degree of correlation between the pre-treatment information and the outcome.

In this case, all designs in studies 1-4 have precision gains relative to the standard design, even when incurring up to 50% sample loss. Study 5 has precision gains over the standard design with up to 30-40% sample loss even though alternative designs did not improve precision at previous correlation levels. Even Study 6 has precision gains over the standard design with up to 20% sample

loss as long as its design combines blocking and pre-post measurement.

More generally, sample loss rarely makes studies 1-4 perform worse than the standard design under the current simulation parameters. Why is this not the case for studies 5 and 6? As Table 5 suggests, Study 5 has the smallest sample size in our sample, making even modest amounts of sample loss consequential and requiring highly correlated variables to make alternative designs attractive.

Study 6 has a larger sample, but the outcome variable we selected is the difference between two aggregate measures that average over four indicators each (see Simas 2022 for details). This results in a continuous outcome centered at 0 with an interquartile range of $[-0.50, 0.25]$. With little variance in the outcome, there is not much room to improve precision through alternative designs, making any sample loss consequential.

7 Guidelines at the Design Stage

While the decision to collect additional pre-treatment information and implement an alternative design needs to be considered on a case-by-case basis, we conclude by enumerating common concerns we believe arise when deciding to implement an alternative design and how we recommend approaching them.

1. *How much explicit sample loss should a researcher expect from an alternative design?* Published articles typically state the experiment's starting sample size and the size of the sample with which the treatment effects are estimated *after* any explicit sample loss. We recommend researchers leverage this information from *recent* experiments in similar contexts to inform expected explicit sample loss rates. For example, the landscape of online survey takers changes rapidly, thus recent experiments conducted on a given platform (e.g., MTurk, Prolific, Lucid, etc.) will provide the best information about expected loss.

   To assist in this, as a part of this article's replication archive, we provide a dataset describing 216 experiments, described in Section 4. A researcher can consult this dataset at the design stage to find examples of experiments similar to their own planned design to reference for

explicit sample loss rates. Our data include whether the experiment implemented pre-post and/or block randomization, the type of experiment (survey, lab, field), what vendor was used to field the experiment (if available), and more.

We have also created an R package called `simprecision` to conduct simulations to estimate precision gains or losses from implementing the standard and alternative designs described in this article. The function allows the researcher to vary many design features, including sample loss incurred from implementing an alternative design. The simulations and visualizations in Section 6 are examples of the package's functionality.

2. *How much implicit sample loss should a researcher expect from an alternative design?* Assessing implicit sample loss is something exclusively available to the individual experimenter since they typically negotiate with survey firms or implementation partners on their own behalf about the cost and length of surveys and interventions. Therefore, individual researchers are best suited to assess how much implicit loss to expect from various design choices. As with explicit sample loss, researchers can use the `simprecision` R package to simulate how varying degrees of implicit sample loss incurred from alternative designs affect precision.

3. *How should a researcher assess the predictiveness of pre-treatment outcome and blocking covariates at the design stage?* While a researcher's initial impulse may be to conduct a pilot study to obtain measures of the predictiveness of pre-treatment variables, we advise against this. If not already cost prohibitive, it possibly decreases the size of the sample they can afford in the full study (a form of implicit sample loss). Moreover, false positive results in pilots with small samples may mislead the researcher into selecting variables that are not highly predictive of the outcome.

   A more promising avenue would be to combine domain expertise with information from previously conducted experiments. For example, in our replication studies, the correlations between pre- and post-treatment outcomes were $r_{DH} = .28$ (quasi measure), $r_{BG} = .70$, and $r_{TH} = .69$. Across six survey experiment replication studies, Clifford, Sheagley, and Piston (2021) find pre-post treatment outcome correlations ranging from $r = .60$ to $r = .90$.

   We do not have comparable evidence for the expected predictiveness of blocking covariates.

We recommend researchers use their own expertise and leverage data from previous experiments if they are concerned about this correlation. In the case of survey experiments, one could also use data from existing public opinion surveys. Finally, `simprecision` allows researchers to vary the predictiveness of pre-treatment outcomes and blocking covariates to see its expected effects on precision, as we demonstrate in Section 6.

4. *Is it a concern that the sample for estimating treatment effects may change under an alternative design?* We find little evidence of this in Section 5.2. However, at the design stage, researchers ought to clearly state their estimand of interest as either a sample average treatment effect (SATE) or a population average treatment effect (PATE) (Hartman 2021). If the estimand is a SATE, the SATE itself may change depending on the design fielded and the sample recruited. However, both the standard and alternative designs will provide for an unbiased estimator of the SATE. If the estimand is a PATE, researchers need to be more cautious to ensure an alternative design does not affect the ability of their sample to generalize to the population.

5. *Is it a concern that alternative designs cause differential post-treatment attrition across treatment arms?* We find little evidence of this in Section 5.3. It is important to note that our results are from online survey experiments where participants are likely accustomed to lengthy batteries and desire to finish the task for compensation. We expect it is more important to consider potential differential attrition across treatment arms caused by an alternative design in field experiments where participants do not have the same incentive structure.

8 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 and pre-post designs. 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 strong implementation of 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, but we hope our findings 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.

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.

Bayram, A Burcu, and Erin R Graham. 2022. "Knowing How to Give: International Organization Funding Knowledge and Public Support for Aid Delivery Channels." *The Journal of Politics* 84 (4): 1885–98.

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, 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.

Curiel, María Ignacia, Cyrus Samii, and Mateo Vásquez-Cortés. 2023. "Democratic Integration of Former Insurgents: Evidence from a Civic Inclusion Campaign in Colombia." *The Journal of Politics* 85 (3): 919–32. https://doi.org/10.1086/723967.

Dietrich, Bryce J, and Matthew Hayes. 2023. "Symbols of the Struggle: Descriptive Representation and Issue-Based Symbolism in US House Speeches." *The Journal of Politics* 85 (4): 1368–84.

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.

Gaines, Brian J, James H Kuklinski, and Paul J Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15 (1): 1–20.

Galasso, Vincenzo, Tommaso Nannicini, and Salvatore Nunnari. 2023. "Positive Spillovers from Negative Campaigning." *American Journal of Political Science* 67 (1): 5–21. https://doi.org/10.1111/ajps.12610.

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.

Goerger, Samantha, Jonathan Mummolo, and Sean J. Westwood. 2023. "Which Police Departments Want Reform? Barriers to Evidence-Based Policymaking." *Journal of Experimental Political Science* 10 (3): 403–12. https://doi.org/10.1017/xps.2022.21.

Harden, Jeffrey J, Anand E Sokhey, and Hannah Wilson. 2019. "Replications in Context: A Framework for Evaluating New Methods in Quantitative Political Science." *Political Analysis* 27 (1): 119–25.

Hartman, Erin. 2021. "Generalizing Experimental Results." *Advances in Experimental Political Science* 385.

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.

Lyon, Nicholas. 2023. "Value Similarity and Norm Change: Null Effects and Backlash to Messaging on Same-Sex Rights in Uganda." *Comparative Political Studies* 56 (5): 694–725. https://doi.org/10.1177/00104140221115173.

Manekin, Devorah, and Tamar Mitts. 2022. "Effective for Whom? Ethnic Identity and Nonviolent Resistance." *American Political Science Review* 116 (1): 161–80. https://doi.org/10.1017/s0003055421000940.

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.

Moore, Ryan T., and Keith Schnakenberg. 2023. *blockTools: Block, Assign, and Diagnose Potential Interference in Randomized Experiments*. https://CRAN.R-project.org/package=blockTools.

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.

Simas, Elizabeth N. 2022. "But Can She Make America Great Again? Threat, Stability, and Support for Female Candidates in the United States." *Political Behavior* 44 (1): 1–21. https://doi.org/10.1007/s11109-020-09607-4.

Tappin, Ben M, and Luke B Hewitt. 2023. "Estimating the Persistence of Party Cue Influence in a Panel Survey Experiment." *Journal of Experimental Political Science* 10 (1): 50–61.