Balancing Precision and Retention in Experimental Design*

Gustavo Diaz† Erin L. Rossiter‡

July 20, 2022

Link to most recent version

Abstract

In experimental political science, researchers can improve the properties of their research design through the choices they make. One important property is the standard error, or the precision, of the estimator. In this paper we consider estimation of the average treatment effect and two common strategies to increase its precision–block randomized and pre-post designs. Holding sample size constant, these design choices can improve precision. In practice, however, implementing these choices may decrease sample size and any gains in precision may be offset. This paper develops guidelines to balance these competing components of precision. We revisit the research design of a published survey experiment on the effect of social media comments on believing misinformation to examine how balancing precision and retention could have influenced this experiment's design decisions. Using simulations, we show how researchers can consider their design's effects on precision at the planning stage. Our findings highlight the importance of considering the practical implications of design choices at the pre-analysis stage to make informed decisions about optimal research design.

*__This paper is a work in progress, please ask before citing.__ We thank Geoffrey Sheagley and participants at MPSA 2022 for helpful feedback.

†Postdoctoral Fellow. Department of Political Science. McMaster University. E-mail: g.diazdr@gmail.com

‡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 political science (Druckman and Green 2021). Tools like DeclareDesign (Blair et al. 2019) and norms like preregistration (Ofosu and Posner 2021) push researchers to carefully consider properties of their research design *before* collecting data. When designing experiments, one goal is to minimize the standard error of the estimated average treatment effects $\text{SE}(\widehat{ATE})$, or the "precision" of the experiment. Because we only get one chance to conduct our randomization, collect data, and generate an estimate of the true ATE, researchers can take advantage of the opportunity to design their experiments to increase their likelihood of recovering an estimate closer to the true ATE.

Researchers consider precision in all parts their research design. The most straightforward strategy to increase precision is increasing sample size if resources permit. Indeed, researchers may allocate a greater proportion of units to the control group if it is less expensive—another strategy to increase precision (e.g., Gerber, Green, and Larimer 2008). Researchers also craft their experimental interventions with precision in mind. Precision suffers if an intervention's effect is highly variable, leading researchers to carefully design their interventions to ensure they have the desired effect (Porter and Velez 2021). As a final example, researchers' care to reduce measurement error in outcomes, perhaps by using an index rather than one item, has a precision motivation (Broockman, Kalla, and Sekhon 2017). Measures with less of noise have less variability. All these strategies and more make an experiment more likely to recover an estimate closer to the true ATE.

While there are many strategies to increase precision, in this paper we focus on two design choices the literature explicitly recommends using with the promise of increased precision. These choices are: (1) using *block randomization* relative to complete or simple randomization and (2) using *pre-post designs* to measure outcomes prior to treatment assignment relative to only measuring the outcome post-treatment. The literature advocates for these

design choices in social science settings to improve 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, 664:103). Regarding repeated measures designs, Clifford, Sheagley, and Piston (2021) recommend researchers to implement this feature whenever possible to improve precision.

Despite the literature recommending block randomized and prepost designs, they are not widely used 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.

Why are these particular design choices not widely used despite the statistical benefits? We believe researchers are cautious about block randomized and pre-post designs because the literature's promise of these designs leading to improvements in precision **assumes sample size is held constant.** However, in practice, researchers often run into contexts where that is unlikely to be the case. For example, if block randomized and pre-post designs require an additional survey wave, participants may drop from the study or the researcher might not be able to afford as many participants, reducing the experiment's sample size. The same decisions that *improve* precision according to advice in the literature might actually *hurt* precision depending on the severity of sample size loss. In this paper, we make the tension between design choices that promise increasing precision but may invite sample loss clear—a tension we suspect researchers are aware of and need guidelines to navigate.

Specifically, we expect researchers consider two ways these two design choices might lead to

[1] 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.

attenuated 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 an pre-treatment survey wave may not be available again for the second wave containing the experimental manipulation. The design choice to add a survey wave would attenuate sample size relative to if the experimental portion of the study were conducted during the first wave. We delineate a second form of sample loss, which we call *implicit sample loss.* This form of attenuation happens when investing in an alternative design forces the researcher to settle with a smaller sample size to begin with, largely 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 her entire budget to only measuring outcomes after delivering treatment.

In this paper, we develop guidelines to navigate the choice to implement block randomized and pre-post designs when a researcher likely faces sample loss from these choices. We investigate the problem from three perspectives to inform our guidelines. First, we reduce the problem to a tractable one-dimensional tradeoff between the competing components of precision. 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 can applied researchers can consider these design choices at the pre-analysis stage.

In sum, we echo the advice in the literature that block randomized and pre-post designs improve precision *if sample size is not affected, either implicitly or explicitly.* The more complicated scenario arises when sample loss might occur. Our simulations critically show that highly predictive blocking covariate(s) and pre-treatment outcome measure(s) can produce precision gains that withstand non-negligible sample loss. We also show that alternative

designs may inadvertently harm precision if sample loss is likely to occur from implementing alternative designs and the incorporated pre-treatment information is not strongly predictive of the outcomes. In any case, we join the conversation on the benefits of simulating experimental designs during the pre-analysis stage (Blair et al. 2019). We highlight how doing so allows a researcher to balance the competing components of precision. Simulations allow a researcher to be confident in whether the precision gains of incorporating pre-treatment information into their design in the form of block randomization and/or pre-post measurement can remain beneficial if sample size is attenutated as a consequence.

This paper 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.

2 Designs to Improve Precision

The most common experimental design implemented in political science has two defining features. First, it assigns treatments using complete randomization, which entails assigning a fixed number of observations across treatment conditions. Second, it measures outcomes only after administering treatments. We refer to a design using complete randomization and post-treatment outcomes measures only as the "standard" design.

Precision motivates researchers to entertain alternative research designs.[2] Researchers can

[2]Other motivations, like features of the research context, may also lead researchers to consider deviations

deviate from the standard design by choosing an alternative randomization procedure or when outcomes are measured. Table 1 maps the alternative designs that we consider across these two dimensions. In this paper, we compare the standard design to three other designs highlighted in Table 1: block randomized, pre-post, and their combination.

We focus on these dimensions for the same reasons that have made these alternative designs well known and advocated for in the political science experimental design literature. First, consider block randomization. Block randomization creates subgroups of units that we expect will respond similarly to the experimental interventions, and randomly assigns treatment *within* these subgroups. This randomization procedure is advantageous because it creates mini-experiments where the treatment and control groups are as similar as possible but for the administered treatment. Block randomized designs can greatly improve precision in social science applications (e.g., Moore 2012) and are advocated for in the literature Pashley and Miratrix (2021a).

What does the literature say about blocking? First, blocking is often not feasible. In some cases one cannot collect pre-treatment covariates before treatment administration. However, Imai et al. advise that *when feasible*, "blocking on potentially confounding covariates should always be used" (Imai, King, and Stuart 2008, 493). Many scholars echo this sentiment Pashley and Miratrix (2021a). Gerber and Green further advise to applied researchers that "[i]n practice, biggest downside" to blocking is incorrectly analyzing the resulting data (Gerber and Green 2012, 79). The idea that blocking rarely has negative consequences in practice assumes sample size is held constant. We draw attention to an unconsidered, important practical implication of blocking – that choosing to block may result in sample loss with potentially offsetting or even negative consequences for precision.

The second dimension of design choices we consider is when to measure outcomes. The

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 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

majority of experiments in political science measure outcomes only post-treatment (Clifford, Sheagley, and Piston 2021) and compare observed outcomes across treatment and control groups to estimate treatment effects. Precision of estimated treatment effects can be greatly improved if pre-treatment measures of the outcomes are also collected and used in one of two ways. First, pre-treatment measures of the outcome can be used to rescale the outcome as a "change score," or the difference between the two measures. 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.

While these design choices have obvious statistical benefits, a critical practical concern often arises when researchers consider implementing them over the standard design— a study may lose sample size as a result. To make these consequences a part of decision-making process when considering alternative designs to improve precision, we next outline how sample size could be attenuated either explicitly or implicitly. First, "explicit sample loss" occurs when the sample is already defined and units who would finish the experiment under the standard design do not finish under an alternative design. Therefore, the sample size was already determined, and blocking or repeated measures led to a smaller final sample. 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 also occurs if the researcher adds many covariates to a pre-treatment survey for blocking or repeated measures

purposes increases survey fatigue and more units drop than would without asking these covariates pre-treatment.

We also draw attention to the scenario where sample loss occurs implicitly. Implicit sample loss happens when investing in an alternative design leads the researcher to settle with a smaller sample size to begin with. Therefore, implicit sample loss is not something one can gather from looking at the 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.

To illustrate the extent of sample size loss in common political science applications, imagine a researcher wishes to conduct a survey experiment in Prolific. In this platform, a five minute survey in a non-representative sample of 1,000 respondents costs USD$1,173. Adding an extra question that takes, in average, two more minutes increases the cost to $1,640. To keep the extra question and stay within budget, the researcher should reduce the sample size to about 720 respondents.[3] For field experiments, an extreme case would imply that, keeping everything else the same, conducting an additional baseline survey to measure pre-treatment outcomes or covariates would double the cost of data collection. With a fixed budget, this translates to half of the sample that a standard design experiment would have.

Depending on its nature, sample loss can also introduce bias. For example, this would be the case if respondent attrition correlates with pre-treatment outcomes or blocking covariates. This is beyond the scope of this paper because we focus on the implications for precision due to sample loss, and these implications persist even when the sample loss does not correlate with outcomes, treatments, or key covariates. In other words, uncorrelated sample size loss is sufficient to illustrate the challenges in optimizing precision when choosing among alternative

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

experimental designs. Moreover, we note two reasons why we set aside the issue of bias. First, several forms of explicit sample loss occur pre-treatment, as when participants drop after the first wave of a panel study where the treatment is randomized in the second wave. Thus we consider scenarios where sample loss is not related to treatment assignment. Sample loss could be related to confounders, like if people attrit from online experiments due to lack of digital literacy (Guess and Munger 2020). However as this example shows, balancing bias along with precision and sample size at the same time requires context-specific statements about the nature and direction of bias that are beyond the scope of this paper. Second, bias 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.

To be sure, block randomization and pre-post designs are useful for reasons other than precision gains. Block randomization can help recover heterogeneous treatment effects by ensuring members of the subgroup are evenly divided between treatment and control conditions. Blocking can also help protect against self-selection bias (King et al. 2007). In addition to precision gains, pre-post designs also provide the opportunity to investigate whether treatment groups are balanced on baseline measures of the outcomes and assess whether any differential attrition across treatment arms may be related to baseline measures of the outcomes. While important objectives, the most common primary goal of experimental design in political science is estimating average treatment effects. Therefore, we set aside other benefits of block randomized and repeated measures designs as secondary to the primary goal of precision of the observed ATE, with a specific focus on design choices that may inadvertently threaten precision despite the literature's promise otherwise.

3 Precision of the Average Treatment Effect Under Alternative Designs

In this section, we first describe the standard research design, including its assumptions, how this design affords researchers an unbiased estimator of the ATE, and the precision of the estimated ATE under this 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 experiment pertains to the random assignment procedure most often used by political scientists called complete randomization. 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. An alternative to complete randomization is simple randomization. In this case, each unit has an independent probability of being assigned to treatment, so that $Z_i \sim \text{Bernoulli}(p)$. Under $p = 0.5$, a sufficiently large sample will tend to have roughly half of the units in each condition. Simple randomization is often used in cases where complete randomization is not feasible, such as when subjects join an experiment on a

rolling basis. For the purposes of this paper, we define the standard experiment as one using complete randomization because complete randomization's guarantee of a fixed number of units assigned to each condition is a desirable property given the small sample sizes often used in social science experiments.

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 paper, we are interested in the average treatment effect as our estimand, as it is the estimand politicial scientists are most commonly interested in: $ATE = E[Y_i(1) - Y_i(0)]$. We can obtain an unbiased estimate of the ATE by calculating the difference in the means observed outcome between 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. Experimental design choices influence $SE(\widehat{ATE})$. We next briefly explain how block randomized and repeated measures designs accomplish this goal.

3.2 Block randomized experiment

One way to decrease $SE(\widehat{ATE})$ is to adjust the randomization procedure from complete to block randomization. Block randomization requires the researcher to 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* in 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, but doing so simplifies the choice of estimator.[4] Because randomization occurs within each block, we can consider each block as 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. This is an unbiased estimator of the ATE in a block randomized experiment. We call this estimator $\widehat{ATE}_{\text{Block}}$ to distinguish a block randomized design from the complete randomized design discussed above.

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

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.[5] Like the $\widehat{ATE}_{\text{Block}}$, the true standard error is a weighted average of within-block standard errors (Gerber and Green 2012, 74):

[4] This is part of a different paper in progress, we will cite when available.

[5] 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.

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

Block randomized experiments yield more precise estimates than the standard design when blocks come from 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, like partisanship, to stratify units. Blocking can also be used with 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 (Moore 2012).

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 is statistical benefit only applies if sample size is held constant. By examining Equation 2, we can see that if the denominator decreases as the numerator decreases, the effects of the pre-post 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 measuring it post-treatment. 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" or as a control variable in the estimation of treatment effects. In what follows, we demonstrate the differencing approach which is more analogous to the true $SE(\widehat{ATE}_{\text{Standard}})$ introduced above.[6]

All assumptions for the standard design and ATE remain the same as in subsection 3.1, but

[6]Deciding between a change score or conditioning approach can affect precision as well, but is beyond the scope of this paper as that decision is a part of the estimation strategy and does not affect sample size attenuation.

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 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. This understanding of the benefits of pre-post designs assumes samples size is held constant. 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 Block randomized, pre-post design

Block randomized and pre-post designs are not mutually exclusive strategies. 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 $SE(\widehat{ATE})$. 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. A researcher can utilize both strategies simultaneously by using pre-treatment information to assign treatment within block and to redefine the outcome (or use covariate adjustment). Indeed, the literature advises that pre-treatment measures of the outcome are often the best covariates as they are usually the most predictive pre-treatment information a researcher has about the potential outcomes.

4 The Precision-Retention Tradeoff

In this section, we demonstrate how block randomized and pre-post designs may be beneficial design choices in terms of increased precision in $\widehat{ATE}$. At the same time, there are costs that accompany these design choices, particularly in terms of sample loss. How do we weigh this tradeoff?

Before considering alternative designs, the simplest alternative to improve precision would be to increase the sample size under the standard design. 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 in the case of Black Americans in the United States (Burge, Wamble, and Cuomo 2020). 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 that the literature promotes as effective ways to increase precision. To see how precision increases, consider the standard error of the $\widehat{ATE}$ in Equation 2. Block randomized and repeated measures designs can reduce $Var(Y_i(0))$ and $Var(Y_i(1))$, but possibly at the cost of reducing N. We draw attention to this specific tension, setting aside the role of the covariance in potential outcomes for now by assuming it is held constant. 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.

To make these competing components of precision more concrete, consider the schedule of potential outcomes for eight units outlined in Table 2. Under the standard design, $Var(Y_i(0)) = 1.25$, $Var(Y_i(1)) = 4.25$, $Cov(Y_i(0), Y_i(1)) = 2.25$, and $N = 8$. Using these inputs to the standard error formula above, $SE(\widehat{ATE}) = 1.19$.

Now consider the precision-retention tradeoff in the case of block randomization. Assume the researcher has good reason to believe units 1-4 and units 5-8 have similar potential outcomes and therefore would make good blocks. We've labeled the observations accordingly. However, assume that in making the choice to use block randomization, the researcher *loses* units, denoted by the rows shaded gray in Table 2.[7] Calculating $SE(\widehat{ATE}_{block})$ standard error will allow us to determine if reducing variation in potential outcomes is worth the loss of sample.

Using the example in Table 2, the inputs to the standard error formula for Block 1 are: $Var(Y_i(0))_1 = .25$, $Var(Y_i(1))_1 = .25$, $Cov(Y_i(0)_1, Y_i(1))_1 = .25$, and $N = 2$. Notice how the variation in potential outcomes *within* the block is much smaller than when considering the

[7] $Var(Y_i(0))$, $Var(Y_i(1))$, and $Cov(Y_i(0), Y_i(1))$ for the $N = 4$ sample are identical to the full $N = 8$ sample so we can compare the effects of sample size loss to gains in precision from block randomizing, all else constant.

Table 2: Schedule of potential outcomes

ID	Block	$Y_i(0)$	$Y_i(1)$
1	1	1	4
2	1	2	5
3	1	1	4
4	1	2	5
5	2	3	8
6	2	4	9
7	2	3	8
8	2	4	9

Note: Rows shaded in gray drop under block randomization.

entire sample. Taken together, $SE(\widehat{ATE}_1) = 1$. Likewise, for Block 2, $Var(Y_i(0))_2 = .25$, $Var(Y_i(1))_2 = .25$, $Cov(Y_i(0)_2, Y_i(1))_2 = .25$, $N = 2$, and $SE(\widehat{ATE}_2) = 1$. Under block randomization with $N = 4$, $SE(\widehat{ATE}) = 0.71$. In this example, even though the sample size is halved, the researcher would rather implement block randomization because the precision gains in doing so outweigh the costs associated with sample loss.

5 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 lead many people to retain the incorrect information, despite the correct information being available. Using the authors' replication archive[8], we imagine how an applied researcher might navigate block randomized and pre-post design decisions.

[8]The Harvard Dataverse replication archive is located here.

Anspach and Carlson's experiment has four arms. The first arm is a baseline condition showing participants a full news article that cited Trump's factually-correct approval rating (36%) from a recent poll. The second arm showed 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 the precision-retention tradeoff 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 (or simple) 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.

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 can't ask how much the respondent's trust a poll before the respondents have been exposed to it. Therefore, the authors did not collect pre-treatment outcomes that we can use to investigate the precision afforded by

alternative designs. 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. In addition to assessing block randomization given the pre-treatment information the resaerchers 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?

5.1 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, we might expect that partisans would have different responses to learning about Trump's approval rating. With this in mind, 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 results as truth for the outcome assessing how much participants trust the poll referenced in the news article or preview. The authors use covariate adjustment and control for several pre-treatment covariates, so we use these reported coefficients when

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

$$\begin{aligned} Y_{i,t2} = 3.28 - .18 * Preview_i - .57 * LiberalComment_i - .62 * ConservativeComment_i \\ +.02 * NFC_i - .01 * NFA_i - .05 * Knowledge_i + .01 * Age_i + .01 * White_i \\ +.02 * Education_i + .01 * Income_i - .45 * Party_i + u_i, \end{aligned} \tag{6}$$

where $\mu \sim N(0, 1)$ 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 data simulated according to Equation 6, we then simulate the different design decisions outlined in Table 1. First, we simulate the standard design 1,000 times with the full sample size. Then, we assess four alternative designs, but penalizing the sample size, assuming only 75% retention due to implementing an alternative design.

The alternative designs are:

1. Complete randomization including the pre-treatment measure of the outcome as a predictor when estimating $\widehat{ATE}$ (Complete + Prepost)
2. Block randomization, blocking only on a three-item indicator of partisanship (Block on PID + Postonly)
3. Blocking on the pre-treatment measure of the outcome and also using it as a predictor when estimating $\widehat{ATE}$ (Block on outcome + Prepost)

4. Blocking on all of the covariates used in Equation 6, including the simulated pre-treatment measure of the outcome, and using a pre-post design (Block on everything + Prepost)

We simulated each of the four 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 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 is only weakly correlated ($\rho = .25$). 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 3. 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.

In sum, this hypothetical example shows that when 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. However, if the blocking covariate 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.

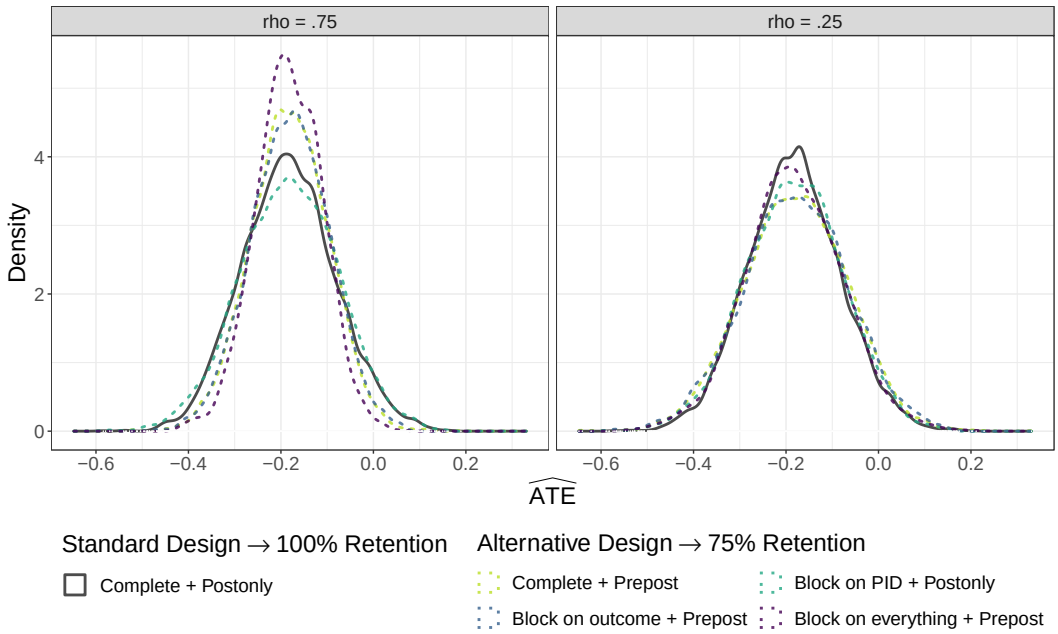

Figure 1: Alternative designs improve precision even with sample attenuation

Table 3: Power of standard design and alternatives with sample loss

	N	Rho=.75	Rho=.25
Complete + Postonly	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

6 Simulation

6.1 Setting

The previous section uses a previously published study to illustrate how researchers can navigate the precision-retention tradeoff. This section uses simulation to show how researchers can navigate the tradeoff at the planning stage. 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 the 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 the trade-off, 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 treatments, potential outcomes, or both. That means our assumption conveys a best-case scenario, yet it is sufficient to illustrate the trade-off.[9]

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 the trade-off 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

[9]We may consider relaxing this assumption in the future.

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 $\alpha = 0.05$

6.2 Results

Figure 3 shows the distribution of statistical power for our two simulation scenarios. Panel A 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

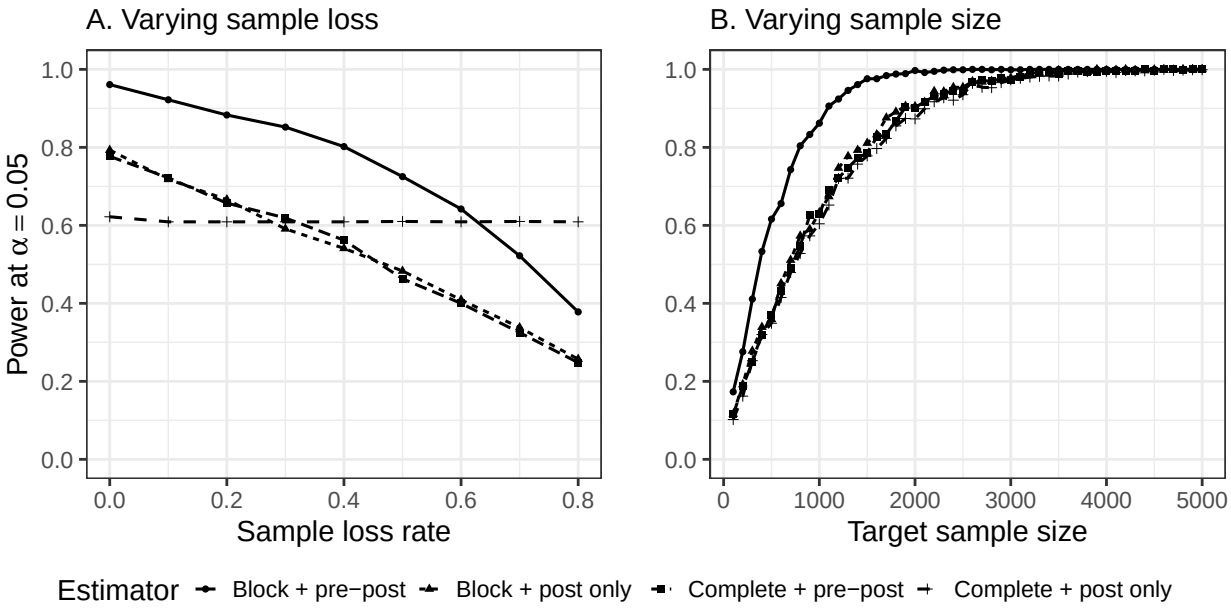

Figure 2: Statistical power for simulated experiments along sample loss rate and target sample size
Note: Each point along the horizontal axis is based on 1,000 simulated experiments.

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

Panel B 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 details of 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 primary source of sample loss comes 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 Panel A, 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 primary source of sample loss comes from being forced to collect a much smaller sample to allow for a multi-wave study, then the horizontal comparisons in Panel A 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, then the comparisons in Panel B provide guideline to determine what would be the minimal target sample to collect under alternative research designs.

7 Conclusion

Previous work proposes deviations from the standard experimental design to improve statistical precision under the assumption that sample size remains constant. This paper develops

standards to choose among alternative designs under explicit or implicit sample loss. This paper advances three important conversations in the political science research design literature.

First, this paper joins others working to shed 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 paper aims to incorporate practical concerns into the public, published conversation on experimental design. We hope this paper 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 paper 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 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. Rather, we hope our findings and guidance lays a path for researchers to understand and consider via simulation the competing components of precision in their experiment.

References

Anspach, Nicolas M, and Taylor N Carlson. 2020. "What to Believe? Social Media Commentary and Belief in Misinformation." *Political Behavior* 42 (3): 697–718.

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

Box, George EP, William H Hunter, Stuart Hunter, et al. 1978. *Statistics for Experimenters*. Vol. 664. 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.

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.

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.

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

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.

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.

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.

Porter, Ethan, and Yamil R Velez. 2021. "Placebo Selection in Survey Experiments: An Agnostic Approach." *Political Analysis*, 1–14.