Balancing Precision and Retention in Experimental Design
Appendix

Contents

A An Example of Balancing Precision and Retention when Blocking

In this Appendix, we use a toy example to illustrate how a block randomized design may be a beneficial design choice in terms of increased precision in $\widehat{ATE}$. Consider the schedule of potential outcomes for eight units outlined in Table B1. 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, $SE(\widehat{ATE}) = 1.19$.

Table A1: 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.

Now consider the tension between improving precision with block randomization in the face of potential sample loss. 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 have 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.[1] Calculating $SE(\widehat{ATE}_{Block})$ will allow us to determine if reducing variation in potential outcomes is worth the loss of sample.

Using the example in Table A1, 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 entire sample. Taken together, $SE(\widehat{ATE}_1) = 1$. Likewise, for Block 2, $Var(Y_i(0))_2 = .25$, $Var(Y_i(1))_2 = .25$,

[1] $Var(Y_i(0))$, $Var(Y_i(1))$, 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.

$Cov(Y_i(0)_2, Y_i(1))_2 = .25$, $N = 2$, and $SE(\widehat{ATE}_2) = 1$. Under block randomization with $N = 4$, $SE(\widehat{ATE}) = 0.71$. In this example, perhaps counterintuitively, 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.

B Estimators for Block-Randomized Experiments

In this Appendix, we discuss alternative estimators for block-randomized experiments. First consider the most common approach in the literature. In a block randomized experiment, the researcher conducts independent experiments in each block and then aggregates their ATE estimates into a single number summary. This aggregation involves computing a weighted average of the estimates across blocks, the main article text describes the block-size weights estimator as the preferred approach in the literature due to its unbiasedness (Humphreys 2009; Gibbons, Serrato, and Urbancic 2018).

However, one could also use precision or harmonic weights (Gerber and Green 2012) according to the following estimator:

$$\widehat{ATE}_{\text{Precision}} = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{h_b} \widehat{ATE}_b. \tag{1}$$

With $h_b = n_b p_b (1 - p_b)$, with p_b as the proportion treated units in block b. As the name suggests, precision weights take into account the proportion of treated units across blocks, whereas block-size weights only consider the size of each block. Using precision weights is equivalent to using block fixed effects or controlling for blocks in OLS regression.

Which weighting scheme is more appropriate? Bowers, Diaz, and Grady (2022) use simulations to argue that the choice of weighting scheme is consequential when the proportions of treated units across blocks correlate with potential outcomes across blocks. In this case, precision weights may lead to biased yet more precise estimates, which may be preferable when the goal is to distinguish an effect from zero. Throughout the main article text, we assume equal proportions of treated units across blocks, so the choice of estimator is trivial.

C Estimators for Pre-Post Designs

The manuscript introduces pre-post designs using the differencing approach, and in this Appendix, we outline an alternative estimator for the ATE. When using a differencing approach, the estimator for the ATE is equivalent to that of a standard design, except that the outcome variable is now the change score is the difference between individual observed outcomes before and after treatment. This is equivalent to using pre-treatment covariates to rescale outcomes, or the difference in differences (Gerber and Green 2012, chap. 4.1).

An alternative approach to analyze data from a pre-post design is to use pre-treatment outcomes as control variables in regression. From this point of view, analyzing experiments with pre-post designs is no different from incorporating covariates in an experiment to enhance precision (Gerber and Green 2012, chap. 4.2; Bowers 2011; Lin 2013).

In this case, the expression $Y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i$ can be used in OLS regression to estimate the average treatment effect β_1 of binary treatment Z_i on outcome Y_i. Controlling for covariate X_i, which in this case corresponds to a vector recording pre-treatment or baseline outcomes. Chapter 4 of Gerber and Green (2012) illustrates the correspondence between the change score and covariate adjustment approaches in pre-post designs.

Using pre-treatment outcomes has two advantages. First, one can control for a proxy of pre-treatment outcomes in cases where measuring pre-treatment outcomes is not feasible. Clifford, Sheagley, and Piston (2021) call this a quasi-pre-post design. In our case, the outcome of interest was how much respondents trusted the result of a poll presented in a survey experimental vignette. This outcome does not make sense before the experimental stimuli is presented, so one could not calculate change scores in this case.

The second advantage of the covariate adjustment approach is that, much like the precision-weighting approach to block randomization, it can yield biased yet more precise estimates of the ATE than change scores (Freedman 2008). This bias comes from the fact that regression adjustment assumes that the pre-treatment outcome is uncorrelated with the error term, whereas the change score estimator does not (Allison 1990). Lin (2013) argues that in most cases the bias is negligible and that using robust standard errors yields asymptotically valid confidence intervals when the conventional OLS standard errors can hurt precision.

D Handcoding Exercise Details

D.1 Sampling

Our hand coding exercise began with collecting the main article text of a sample of experiments in political science journals published in 2022 or 2023. We collected articles from the following six journals: *American Political Science Review* (APSR), *American Journal of Political Science* (AJPS), *Journal of Politics* (JOP), *Political Behavior* (PB), *Comparative Political Studies* (CPS), and *Journal of Political Science* (JEPS).

We chose this set of journals for several reasons described in the article. We also discussed broadly how we determined whether an article was from our population of interest in the article text, but we describe those criteria more specifically here. First, two research assistants (RA) visited the journal website and reviewed the meta-data for every published article in 2022 and 2023. First, to belong to our population of interest, the article must have been a "full-length" publication for the journal. We excluded short format articles ("Letters" from APSR and "Short Reports" from JEPS) out of the concern that authors would exclude the details we are most interested

in from the main text in contexts with strict word count limits. This constraint should not preclude detailed explanation of design decisions in full-length articles. Second, the RA read the title and abstract for any mention of an experiment. At this step, the RAs were instructed to err on the side of caution and include the article even if they were not sure it met all of the criteria. If the title or the abstract did not mention an experiment, the article was not included, therefore we may have some false negatives. This procedure identified a tentative list of articles from our population. To ultimately be included in our sample, an article must meet several other criteria that we identified from reading the full text of the article.

- The article must contain an original randomized experiment, meaning the authors randomized some intervention and collected new data.This excludes methods papers that reanalyze another article's replication archive and quasi-experimental designs.
- We also exclude experiments designed to demonstrate the contribution of new methods, for example, reducing social desirability bias in survey research. We exclude these articles because we are interested in the design decisions of applied researchers seeking to make a substantive contribution.
- We exclude conjoint experiments because there is not, to our knowledge, methodological work that incorporates block randomization in this context with extremely high-dimensional combinations of randomized features. Relatedly, including conjoint experiments would inflate the number of experimental conditions unfairly in our descriptive analyses.

A total of 227 articles with 336 unique experiments met our criteria. The next step was to hand code several features from a sample of the experiments.

D.2 Hand coding

We took a sample of approximately 50% of the articles identified as applied experimental work in the six journals published in 2022-2023. This sample included 121 articles with 216 unique experiments.

We then read the methods, results, and discussion sections to hand code for several concepts. We only reviewed the article text, including footnotes. We did not review any appendices or preanalysis plans for this information. We do not review beyond the main text because we aim to describe how precision and retention decisions are made, but also importantly, how they are discussed in peer-reviewed, published articles.

We code for the following concepts:

- **Experiment identifier**
  - If an article contains more than one original experiment, each experiment is given a unique identifier determined by the order the experiments were discussed. Each experiment is given a unique row in the dataset and coded separately.
  - Note: Often, separate experiments or randomizations were conducted, but the main analyses pool the samples. In these cases, we code for the features of the pooled design.
- **Discussed balancing precision and retention**
  - Code "1" when the article article discusses balancing precision and retention as a relevant concern for the experiment. This means the article discusses alternative research designs in terms of precision, like trade-offs, pros, or cons involved with different sample sizes, different randomization procedures, etc. *in terms of precision.* Examples:
    - "We used covariates in estimation of treatment effects *to increase precision.*" (could be used for other reasons)
    - "We used block randomization *to increase precision.*" (could be used for other reasons)
    - "We did X in our design to avoid loss to sample size."
    - "We did X in our design to have the biggest sample size possible within the budget," like not include a pre-treatment wave, only ask a limited number of questions in the pre or post-treatment survey, exclude a certain experimental condition, etc.
    - "We used a pure control to minimize variation in the potential outcomes." (could be used for other reasons)
    - "We did not include a baseline survey because we wanted to collect larger sample"
  - Code "0" if the article does not discuss their navigation of balancing precision and retention in the experiment. Importantly, similar design decisions we code as a "1" above may be discussed *without* considering the balance of precision and retention. Examples:
    - "We did not collect pre-treatment measures of the outcome because we did not want to prime the topic before treatment."
    - "Our budget afforded a sample size of X."
    - "We conducted a power analyses, and we are powered to detect effects with X sample size." (Without further entertainment of alternative research designs, like alternative sample sizes.)
- **Pre-treatment covariates**
  - Code "0" if the design did not have pretreatment covariates or it was not possible to infer that it did.
  - Code "1" if the design did collect or was provided pretreatment covariates
  - Note: We were liberal in our coding of this variable. Articles often do not mention *when* covariates are collected. If we could determine that covariates were measured even if the article does not explicitly state that they were measured, we coded it as "1". For example, articles often mention balance tests located in the appendix without mentioning the collection of these variables, so we assume pretreatment covariates were collected. We also code as "1" if covariates were collected after treatment administration.

- **Block randomization**
  - Code "1" if the article mentioned that the experiment used block randomization.
  - Code "0" if the article did not mention whether the experiment used block randomization.
- **Pre-post**
  - Code "1" if the article mentioned that the experiment used prepost estimation.
  - Code "0" if the article did not mention whether the experiment used prepost estimation.
- **Sample size**
  - Sample size in the analysis estimating main treatment effects.
  - Note: When in doubt, we err on the side of the larger sample size. Sometimes the sample size used in estimation is not available in the main article text (e.g., analysis excludes inattentive respondents but that sample size is not reported), so we code the sample size that is reported (e.g., the full sample), again erring on reporting the larger sample size when it is unclear what the sample size was for the main analysis.
- **Number of experimental conditions**
  - Number of unique treatment arms in the experiment.
  - Note: Sometimes articles mention randomizations that are not included in the main analyses. For example, if topics are randomized in addition to the main randomization of interest, and the analyses always pool across topics, we do not code the topics as unique treatment arms.

E Replication Study Details

In this Appendix, we provide additional details on our replication study of Dietrich and Hayes (DH), Bayram and Graham (BG), and Tappin and Hewitt (TH). This research was deemed exempt by a university IRB. It adheres to APSA's Principles and Guidance on human subjects research. We obtained informed and voluntary consent by informing all survey participants about they task and its compensation prior to them beginning it. After reading our informed consent document, consent was documented by participants answering "I consent" to an online survey question.

E.1 Cost

The total cost to conduct these three replications was $7487.17 ($1,222.00 for DH, $2820.00 for BG, and $3,445.17 for all three waves of TH). Costs come only from the payments to participants and the associated CloudResearch Connect fees (25% of the payment to participants). Please see our preregistration for the participant compensation associated with each survey.

E.2 Applying exclusion criteria

Before applying our exclusion criteria, we recruited a total of 1334 participants to the DH replication, 3050 participants to the BG replication, and 2611 participants to the TH replication. We preregistered excluding participants who failed the attention check included in all three replication studies (29 in DH, 26 in BG, and 28 in TH). We did not preregister excluding any other participants. However, after fielding the studies, we realized two additional exclusion criteria were needed. First, for the DH replication, we set a quota in CloudResearch Connect to only recruit participants who were Black or African American. Some participants (23, less than 2%) in the DH replication did not indicate Black or African American to our race and ethnicity survey question, so we exclude them as they are not members of study's population of interest. Second, for TH replication, we made a mistake in the implementation of the replication. In their design, participants without a partisan affiliation were asked a forced-choice question about which party they would prefer to be in power in order to impute a partisan affiliation. Partisan affiliation is needed in TH's measurement of the dependent variable, which is level of agreement with one's inparty position. We did not include the forced-choice question. Therefore, we will need to exclude these participants from our analyses. We apply this exclusion prior to assessing any sample loss, as it is more similar to a screener/qualifier question for the study to only include members of the population of interest. When doing so, we exclude 274 participants (11%). Including non-partisans or not is a substantively interesting decision. TH actually estimate their effects under both scenarios, saying "We also analyze the data excluding Independents in the Appendix; the results are the same" (pg 54). Our decision is bolstered by the fact that TH find consistent results excluding Independents.

After applying our exclusion criteria, we have the following sample sizes: $N_{DH} = 1282$, $N_{BG} = 3024$, $N_{TH} = 2299$.

E.3 Assessing explicit sample loss

Table E2 shows the rate at which observations were lost—what we consider explicit sample loss—across the three alternative design's per study. We first list the starting sample size for each alternative design. Within a replication study, participants were randomized to a design, therefore these sample sizes differ only due to expected variability from this randomization.

Our key interests are: (1) how much sample loss occurred per design and (2) whether the amount of loss differed across designs. First, we assess DH and BG–both single-wave studies. Both studies feature little sample loss regardless of alternative design. Only 2-3 and 5-7 observations per design did not finish the study in DH and BG, respectively. And the proportion of these units that dropped *post-treatment* is even smaller. We care particularly about post-treatment sample loss because it compromises the estimation of unbiased treatment effects if units attrit from the study due for reasons correlated with their treatment assignment. The final column of Table E2

shows the p-value associated with Fisher's exact tests of the null hypothesis of no difference in proportion comparing post-treatment sample loss between each study's standard design and the two alternative designs. In both the DH and BG replications, we fail to conclude there is any difference in post-treatment sample loss between the standard design and either alternative design. We conclude that post-treatment sample loss is of little concern in these single-wave survey experiments. All three designs asked the same eight demographic pre-treatment items. To implement the alternative designs, DH asked an additional three pre-treatment questions and BG asked an additional four pre-treatment questions. While seemingly a small number of questions, these questions led to a sizable increase the survey length for these short surveys. Encouragingly, the additional questions and survey length caused by the alternative designs did not invite a distinguishable amount of additional sample loss relative to the standard design. In such contexts, this replication exercise provides evidence that can alleviate concern over any explicit sample loss that might occur due to implementing a pre-post and/or block randomized design.

Before turning to the results in Table E2 for the TH replication, which allows us to assess explicit sample loss in a multi-wave context, we give an overview of the design elements that occured in each wave. Wave 1 randomized participants to a design and collected needed pre-treatment information. The standard design collected only the the 6 demographic items. The alternative designs also collected an additional 12 items (pre-treatment measures of the participants' attitudes on the 5 topics of interest, 6 feeling thermometers, and a measure of ideology), more than doubling the length of this survey for those assigned to an alternative design. Between Waves 1 and 2, we used multivariate continuous block randomization to randomize treatment to participants assigned to this design. Participants assigned to the standard design or the pre-post only design were randomized to treatment when they entered the Wave 2 study. Then, participants were exposed to their randomized treatment of party cues at the issue level and whether or not they would answer the five attitude items. Wave 3 was a final wave in which *all* participants answered the attitude items.

Next we turn to the results in Table E2 for the TH replication. Because respondents had to return to take three waves of surveys across 1-1.5 weeks, it follows that TH has more total sample loss (losing units at any point in the process) than the DH and BG as single-wave experiments—13.0%, 14.2%, and 20.7% of units dropped at some point in the process across the standard, pre-post, and pre-post with blocking designs, respectively. It is unsurprising that nearly all of these units in the standard and pre-post design dropped *after* treatment assignment.[2] Wave 1 was a short survey, and as we observed with the DH and BG single-wave studies, few units drop during a wave after they have consented. Patterns differ with the block-randomized design. There is relatively more sample loss overall, but the amount of that loss that occurs post-treatment is indistinguishable from the standard design. The reason that there is more pre-treatment sample loss in the block-randomized design is because we implemented multivariate continuous blocking, utilizing several pre-treatment covariates. We decided to exclude any observation from the block randomized treatment assignment, and thus from participating in the experimental portion of the study, if they had any missingness in these covariates. This decision excluded 58 observations, which accounts for the discrepancy between a this design featuring a higher overall sample loss than the others, but a similar rate of post-treatment sample loss. We conclude that implementing pre-post and/or block randomized designs does 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, they may incur additional explicit sample loss. This sample loss is pre-treatment, thus the researcher is not risking biased estimation of their treatment effects by using block randomization relative to the others designs assess in our replication exercise. Nevertheless, our exercise demonstrates that, if missingness in pre-treatment covariates used in blocking is nonrandom, there may be systematic difference between the *sample* a researcher ultimately uses to estimate their treatment effects when using block randomization relative to another design. We assess this question in the next subsection (Appendix Section E.4. Then, we discuss the separate question of whether there is differential post-treatment attrition across treatment arms depending on the design in use. These results are in Appendix Section E.5.

Table E2: Explicit Sample Loss by Study and Alternative Design

Study	Design	Starting sample size	All loss	Post-treatment loss	p-value
Dietrich & Hayes	1. Standard	426	0.005 (2)	0.000 (0)	
	2. Pre-post	427	0.007 (3)	0.005 (2)	0.499
	3. Pre-post & blocking	429	0.007 (3)	0.002 (1)	1
Bayram & Graham	1. Standard	1008	0.005 (5)	0.001 (1)	
	2. Pre-post	1010	0.005 (5)	0.002 (2)	1
	3. Pre-post & blocking	1006	0.007 (7)	0.005 (5)	0.124
Tappin & Hewitt	1. Standard	752	0.130 (98)	0.125 (94)	
	2. Pre-post	780	0.142 (111)	0.141 (110)	0.368
	3. Pre-post & blocking	767	0.207 (159)	0.132 (101)	0.702

Note: Descriptive statistics for explicit sample loss across replication studies. The starting sample size column shows the number of observations randomized to each alternative design. The next two columns show the proportion and count in parentheses of how many observations were lost in total and specifically post-treatment. The final column shows p-values from Fisher's exact tests of the null hypothesis of no difference in proportion comparing post-treatment sample loss between each study's standard design and the two alternative designs.

[2]For the standard design and the pre-post only design, we define post-treatment attrition as a unit dropping from the study anytime during Waves 2 or 3 or not returning for Wave 3. The block-randomized alternative design has a slightly different definition because treatment was assigned *prior* to Wave 2 beginning, so we also deem an observation as attriting post-treatment if they do not return for Wave 2.

E.4 Differential sample inclusion

It is a concern 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. In this Appendix, we use our replication studies to provide evidence for this question. We find little to no evidence that samples resulting from alternative designs differ in meaningful ways. First we note that sample inclusion is a different concern from *post-treatment attrition*, which we address in the next appendix. Sample inclusion asks the question of whether the sample with which you randomize treatment and estimate treatment effects differs, thus resulting in different target SATEs (sample average treatment effect). Thus, bias is not a concern in the discussion of sample inclusion, but it will be the primary concern when considering post-treatment attrition next in Appendix Section E.5.

Table E2 shows that sample loss is trivially small in the first two studies we replicate. Therefore, in these two cases, we conclude that implementing alternative designs–which requires asking additional political pre-treatment items–do not differ in the samples used to estimate average treatment effects. However, for the Tappin and Hewitt (TH) replication, we find sizable sample loss throughout the three-wave study. This provides us the opportunity to understand whether implementing alternative designs that require many additional pre-treatment items (in this case, all political question items) results in different samples that are used to estimate the average treatment effect.

Specifically, in the TH replication, we ask the same 6 demographic items in all three designs. The alternative designs also collected an additional 12 items (pre-treatment measures of the participants' attitudes on the 5 topics of interest, 6 feeling thermometers, and a measure of ideology). These items more than doubled the length of this survey for those assigned to an alternative design. They also presented respondents in the alternative design with many political questions, whereas those in the standard design only had the standard branching two-part partisanship question (identification and strength of identification or which party one leans to if indicating Independent).

We use the TH replication to explain the concern of different samples arising from different designs, starting with three main reasons this concern arises in this context:

1. First, as discussed in Appendix Section E.3, the alternative designs in this study asked over twice as many pre-treatment items as the standard design. Therefore, survey fatigue could cause units to drop from the alternative designs in ways that would make the samples different, such as if fatigue was related to education, age, etc. This is a common concern when implementing many additional pre-treatment items.
2. Second, all of the additional items in the alternative designs pertained to political attitudes. This represents the kind of items researchers implementing a pre-post and/or block randomized design would likely need to include in their pre-treatment battery. However, 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 pre-treatment. For instance, people who have a strong aversion or disinterest in politics may drop from the study.
3. Third, because we implemented multivariate continuous blocking utilizing several pre-treatment covariates in the TH replication, we excluded any observation from the block randomized treatment assignment if they had any missingness in these covariates. As discussed in Appendix Section E.3, this decision excluded 58 observations. If missingness is related to characteristics, such as an aversion to politics, that is another way implementing this particular alternative design could affect sample composition.

Next we turn to evidence from the TH study. First, we assess whether sample inclusion across the three experimental designs is related to demographics we collected. Because these items are not continuous (age, gender, race/ethnicity, income, and partisanship), we create binary indicators for each response option. This results in 24 characteristics we can examine. Table E3 shows results regressing each characteristic on an indicator for the pre-post alternative design and the pre-post and blocking alternative design, therefore the omitted category is the standard design. The sample size for these models is all units who did not drop at any time in the study (the starting sample size minus all losses in Table E2). We find no evidence of difference between the sample resulting from the pre-post design and the sample resulting from the standard design (row one of the table). All measured demographic characteristics are statistically indistinguishable between the design samples. We find one instance where the difference in sample between pre-post and block randomization differs from the standard design. The sample resulting from the pre-post and blocking design had fewer participants who "leaned" toward the Republican party (9.5% of the participants in the standard design and 6.4% in the pre-post and block randomization alternative design, $p = 0.044$). We can also compare the two alternative designs' samples to each other. The final row of Table E3 provides p-values comparing the pre-post and pre-post with block randomization alternative designs. We find no evidence across the 24 models that the samples resulting from the two alternative designs differ.

Second, we assess whether sample inclusion between the two alternative designs is related to political measures we collected. (Recall we do not ask these items in the standard design in order to assess the effect of including them in the alternative designs. See Appendix Section E.3.) We measured ideology using a five-point scale, and assess sample inclusion for each response option separately. All other items (six feeling thermometers measured on a 0-100 scale and five pre-treatment policy questions measured using a five-point Likert scale) were considered continuous. This resulted in 18 political measures. Table E4 presents results. We find no evidence that sample inclusion was related to any of these items.

In sum, while it is a concern 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, we find very little evidence in support of this concern across three replication studies. However, we note that this evidence is limited as we can only assess differential sample inclusion based on the pre-treatment covariates we observed, and the number of items we could include was limited to for budgetary reasons. Nevertheless, this original data collection provides reassurance that sample loss from alternative designs does not result in meaningfully different samples.

Table E3: Assessing Differential Inclusion in the Sample Across Designs–Pre-treatment Demographics Asked in All Designs

	Age (18-24)	Age (25-34)	Age (35-44)	Age (45-54)	Age (55-64)	Age (65-74)	Age (75-84)	Gender (Man)	Gender (Woman)	Race/Ethnicity (White)	Race/Ethnicity (Black or African American)	Race/Ethnicity (Asian)	Race/Ethnicity (Hispanic or Latino)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Pre-post	−0.007	−0.006	−0.011	0.004	0.013	0.004	0.003	0.045	−0.045	−0.021	0.015	0.001	0.008
	(0.017)	(0.025)	(0.025)	(0.020)	(0.016)	(0.011)	(0.004)	(0.027)	(0.027)	(0.024)	(0.019)	(0.014)	(0.018)
Pre-post & blocking	−0.007	−0.023	0.035	−0.017	0.005	0.005	0.002	0.039	−0.036	−0.005	0.015	−0.021	−0.014
	(0.017)	(0.026)	(0.025)	(0.021)	(0.017)	(0.012)	(0.004)	(0.028)	(0.028)	(0.024)	(0.019)	(0.014)	(0.018)
Intercept	0.107*	0.321*	0.272*	0.162*	0.092*	0.043*	0.003	0.471*	0.526*	0.757*	0.128*	0.075*	0.124*
	(0.012)	(0.018)	(0.018)	(0.014)	(0.012)	(0.008)	(0.003)	(0.020)	(0.020)	(0.017)	(0.014)	(0.010)	(0.013)
Observations	1,931	1,931	1,931	1,931	1,931	1,931	1,931	1,931	1,931	1,931	1,931	1,931	1,931
F-test p-value Pre-post = Pre-post & blocking	0.992	0.496	0.068	0.300	0.649	0.907	0.784	0.835	0.753	0.505	0.983	0.122	0.245

	Inc (<20K)	Inc (30-34K)	Inc (35-49K)	Inc (50-74K)	Inc (75-99K)	Inc (100K+)	Party (Strong Dem)	Party (Weak Dem)	Party (Lean Dem)	Party (Lean Rep)	Party (Weak Rep)	Party (Strong Rep)
	(14)	(15)	(16)	(17)	(18)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
Pre-post	0.002	0.010	0.015	−0.010	0.024	−0.042	0.016	0.023	−0.018	−0.017	−0.005	0.001
	(0.016)	(0.018)	(0.019)	(0.023)	(0.020)	(0.024)	(0.025)	(0.024)	(0.019)	(0.015)	(0.019)	(0.016)
Pre-post & blocking	−0.022	−0.014	0.024	0.007	0.029	−0.024	−0.005	0.023	−0.014	−0.031*	0.021	0.007
	(0.016)	(0.018)	(0.019)	(0.024)	(0.021)	(0.025)	(0.025)	(0.025)	(0.020)	(0.015)	(0.020)	(0.016)
Intercept	0.098*	0.118*	0.125*	0.240*	0.144*	0.275*	0.280*	0.239*	0.157*	0.095*	0.141*	0.089*
	(0.011)	(0.013)	(0.014)	(0.017)	(0.014)	(0.017)	(0.018)	(0.017)	(0.014)	(0.011)	(0.014)	(0.011)
Observations	1,930	1,930	1,930	1,930	1,930	1,930	1,931	1,931	1,931	1,931	1,931	1,931
F-test p-value Pre-post = Pre-post & blocking	0.128	0.189	0.644	0.499	0.807	0.458	0.400	0.998	0.837	0.369	0.203	0.724

Note: *p<0.05; **p<[0.**]; ***p<[0.***]

Table E4: Assessing Differential Inclusion in the Sample Across Designs–Additional Pre-treatment Political Questions Asked in Alternative Designs Only

	Ideology (Very Lib.)	Ideology (Lib.)	Ideology (Somewhat Lib.)	Ideology (Moderate)	Ideology (Somewhat Cons.)	Ideology (Cons.)	Ideology (Very Cons.)	Feeling Thermom. (Voters - Reps)	Feeling Thermom. (Voters - Dems)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Pre-post & blocking	−0.001	0.018	−0.035	−0.004	0.014	−0.006	0.014	1.808	0.001
	(0.021)	(0.025)	(0.021)	(0.019)	(0.018)	(0.017)	(0.012)	(1.614)	(1.474)
Intercept	0.163*	0.267*	0.180*	0.130*	0.108*	0.106*	0.045*	39.720*	58.900*
	(0.014)	(0.017)	(0.014)	(0.013)	(0.012)	(0.012)	(0.009)	(1.117)	(1.017)
Observations	1,275	1,275	1,275	1,275	1,275	1,275	1,275	1,269	1,277

	Feeling Thermom. (Elites - Trump)	Feeling Thermom. (Elites - Obama)	Feeling Thermom. (Elites - Biden)	Feeling Thermom. (Elites - Romney)	Salestax Pref.	Pension Pref.	Fed. Audit Pref.	Foreign Aid Pref.	Healthcare Pref.
	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
Pre-post & blocking	0.614	−0.696	−2.794	−0.245	0.066	−0.040	−0.107	0.127	0.108
	(1.935)	(1.843)	(1.804)	(1.324)	(0.095)	(0.089)	(0.092)	(0.097)	(0.088)
Intercept	26.972*	62.657*	47.361*	31.133*	3.851*	4.449*	4.018*	4.099*	4.821*
	(1.350)	(1.273)	(1.248)	(0.916)	(0.066)	(0.062)	(0.063)	(0.067)	(0.061)
Observations	1,249	1,275	1,270	1,271	1,277	1,276	1,277	1,277	1,274

Note: *p<0.05; **p<[0.**]; ***p<[0.***]

E.5 Differential attrition

Next we turn to examining post-treatment attrition. There are two separate concerns here. First, are attrition patterns different *across designs*? Second, are attrition patterns different *across treatment arms* across designs? The latter may bias treatment effect estimation. As in Appendix Section E.4, we only investigate attrition in the TH replication since attrition was trivially small in the other replications. Specifically, we investigate whether *any* sample loss across designs is related to the pre-treatment measures we collected. The pre-treatment characteristics we examine are the same as in Section E.4. In the following analyses, our sample of interest is anyone randomized to one of our three designs of interest (standard design, pre-post design, or pre-post and block randomized design). Our outcome of interest is an indicator for whether the individual attrited post-treatment (1) or not (0).

E.5.1 Post-treatment attrition patterns by design

In Figure E1, we present results from regressing the post-treatment attrition indicator on an indicator for each alternative design, the an indicator for the pre-treatment demographic characteristic, and an interaction between alternative design indicator and pre-treatment characteristic. We plot the p-values from the two interactions of interest. The p-values tell us whether the a pre-treatment characteristic's effect on post-treatment attrition is different between an alternative design and the standard design. We find no evidence of differential attrition by design across the 24 pre-treatment demographic characteristics measured. While it appears men attrited less in the pre-post design than in the standard design, this interaction did not reach conventional levels of statistical significance ($p = 0.057$).

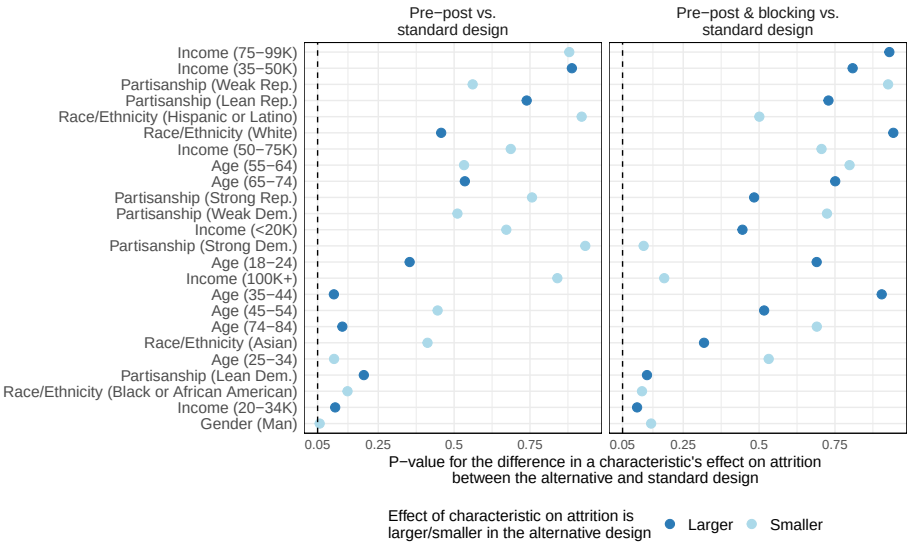


Figure E1: Assessing Differential Post-Treatment Attrition by Design–Demographics Asked in All Designs

In Figure E2, we conduct a similar exercise for all measure political characteristics. As described in Appendix Section E.4, these questions were only asked in the alternative designs, thus we cannot assess if attrition is larger/smaller based on these characteristics in alternative designs, relative to the standard design. Instead, Figure E2 compares the two alternative designs to each other. Again we find no evidence of differential attrition by design across the 19 pre-treatment political characteristics measured. While it appears stronger agreement that the Federal Reserve Bank should be audited by Congress led people to attrit more in the pre-post and block randomized design design than the pre-post only design, this interaction did not reach conventional levels of statistical significance ($p = 0.059$). Encouragingly, we find no evidence of differential post-treatment attrition across alternative designs in this replication exercise.

E.5.2 Post-treatment attrition patterns design *and* treatment arm

Next, we assess the more concerning patterns of differential attrition. If *any* design causes different attrition between treatment and control, this may bias treatment effect estimates. Here, we assess if alternative designs invoke differential attrition across treatment arms differently than the standard design.

Again, our outcome is post-treatment attrition. We now regress the post-treatment attrition indicator on an indicator for each alternative design, an indicator for treatment assignment, the an indicator for the pre-treatment demographic characteristic, and the interaction of these variables. We plot the p-values from the two three-way interactions of interest (interactive effect of a given alternative design, treatment status, and presence of covariate on attrition).

The p-values tell us whether the a pre-treatment characteristic's effect on post-treatment attrition *across treatment arms* is different between an alternative design and the standard design. We find some, but little, evidence of differential attrition by design across the 24 pre-treatment demographic characteristics measured in Figure E3. Asian participants attrited *less* from treatment in the pre-post design relative to the standard design. Black and African American respondents attrited *more* from treatment in this same comparison of designs. Moving to the block randomized design, 65-74 year old participants and weak Democrats attrited less from treatment than in the standard design. Turning to political characteristics only measured in the alternative designs, we find that liberal participants attrited less from treatment in the block randomized design relative to the pre-post design.

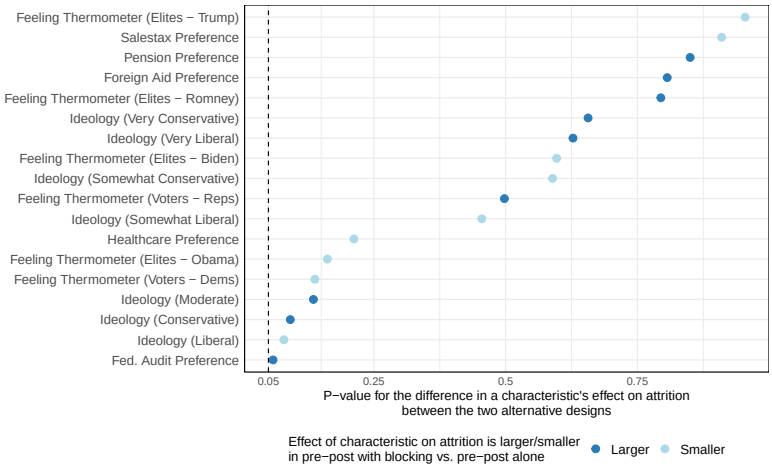


Figure E2: Assessing Differential Post-Treatment Attrition by Design–Political Variables Asked in Alternative Designs Only

In sum, we find that the majority of measured covariates do not explain differential attrition across treatment arms by design. We do find some, but little, evidence of patterns of attrition that correlate with treatment assignment. To be sure, these patterns may bias treatment effect estimates. However, it is unclear if these patterns are driven by the design, noise, or something else. In particular, when attrition patterns differ between the standard and alternative designs, 3 out of 4 times the patterns is that participants attrited *less* in the alternative design (Figure E3). This suggests that the alternative designs are not significantly contributing to higher attrition rates in ways we can explain with the data at hand, but that other factors might be influencing participant dropout.

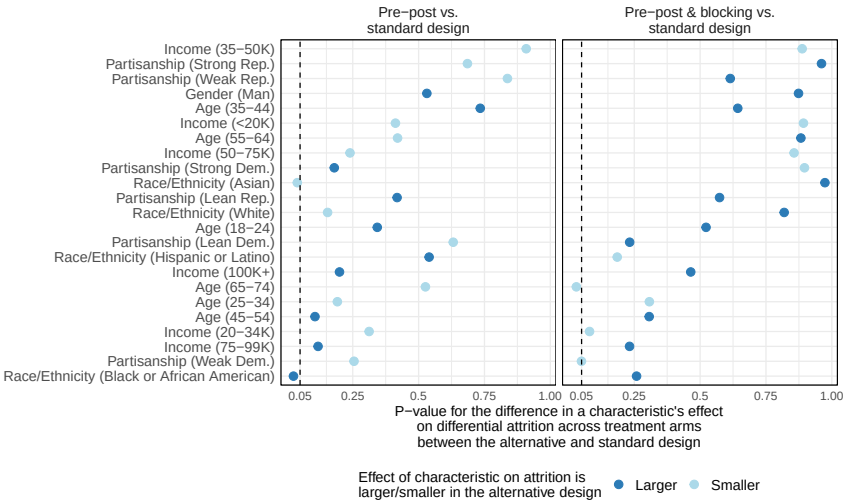


Figure E3: Assessing Differential Post-Treatment Attrition Across Treatment Arms by Design–Demographics Asked in All Designs

E.6 Treatment Effects

Tables E5, E6, and Table E7 provide treatment effect estimates for the three replication studies. Each table provides four estimates, one using observations assigned to the standard design (column 2), one using observations assigned to the pre-post design (column 3), one using observations assigned to the pre-post and block randomized design (column 4), and one pooling all observations in the replication (column 1). In DH, we fail to replicate the original study's treatment effects. Three of the four estimations result in a treatment effect that we cannot distinguish from zero. The pre-post estimated treatment effect is distinguishable from zero, but it contains the opposite sign as DH's experimental findings. In BG, we similarly fail to replicate the original study's treatment effects. We cannot distinguish any of the estimates from zero. Finally, in the TH replication, we replicate the original study's treatment effects in the block randomized study only. The other estimations are not distinguishable from zero.

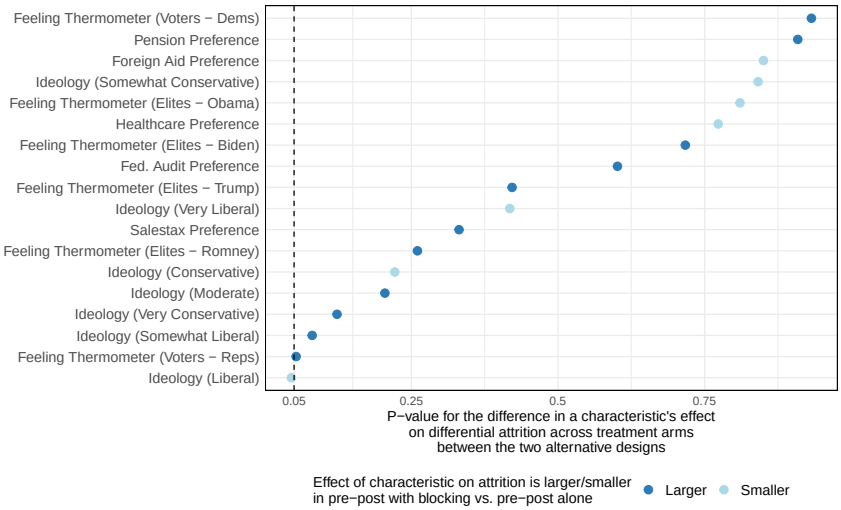


Figure E4: Assessing Differential Post-Treatment Attrition Across Treatment Arms by Design–Political Variables Asked in Alternative Designs Only

Table E5: Treatment Effect Estimates By Design, Dietrich and Hayes Replication

	Pooled	Standard design	Pre-post	Block randomized
Treatment Effect	−0.023	−0.022	−0.058*	0.011
	(0.015)	(0.023)	(0.027)	(0.023)
Pre-treatment outcome measure			0.461*	
			(0.071)	
Intercept	0.893*	0.907*	0.523*	
	(0.010)	(0.015)	(0.064)	
Observations	630	212	209	209
Block fixed effects				Yes

* $p < 0.05$

Table E6: Treatment Effect Estimates By Design, Bayram and Graham Replication

	Pooled	Standard design	Pre-post	Block randomized
Treatment Effect	0.024	0.033	0.018	0.011
	(0.028)	(0.049)	(0.033)	(0.032)
Pre-treatment outcome measure			0.748*	
			(0.034)	
Intercept	0.410*	0.392*	0.124*	
	(0.020)	(0.035)	(0.023)	
Observations	1203	399	402	402
Block fixed effects				Yes

* $p < 0.05$

Table E7: Treatment Effect Estimates By Design, Tappin and Hewitt Replication

	Pooled	Standard design	Pre-post	Block randomized
Treatment Effect	0.022	0.029	0.003	0.051*
	(0.012)	(0.022)	(0.015)	(0.017)
Pre-treatment outcome measure			0.706*	0.716*
			(0.032)	(0.060)
Intercept	0.579*	0.556*	0.184*	
	(0.009)	(0.015)	(0.022)	
Observations	1936	656	669	610
Block fixed effects				Yes

* $p < 0.05$

E.7 Implementing Block Randomization in Practice

It is beyond the scope of this article to explain how to implement block randomization in every context. However, we acknowledge that practical issues like implementation can be a major hurdle to implementing beneficial designs. Therefore, in our replication archive, we provide multiple resources for assisting researchers new to implementing block randomization. First, we provide the Qualtrics Survey File (.qsf) for these three replicated experiments. This allows a researcher with access to Qualtrics to upload the exact implementation of each experiment we fielded, importantly including the implementation of block randomization. Two of the examples (DH and BG) block randomize within a single-wave experiment, a common context for survey experiments. The final example (TH) implements multivariate continuous blocking in a multi-wave context. The block randomization occurs *outside* the Qualtrics environment in an R script (also provided in the repliation archive) and is uploaded by the researcher between pre-treatment and experimental waves. While more complicated, we intend for a researcher to utilize our examples to confidently implement block randomized experiments if they conclude doing so will produce important precision gains.

F Simulation Study Details

F.1 Deviations from pre-registration

At the pre-analysis stage, we designated variables in the data as blocking covariates or pre-treatment outcome proxies. Our selection was based exclusively on the discussion presented in the original articles and supplemental materials. We only consulted the replication data to determine whether the variables we wished to use were available.

Since we did not look at numerical relationships between these variables and the outcome of interest at the pre-analysis stage, we did not anticipate how many of them would exhibit low correlation with the outcome of interest. This made them poor candidates for a pre-post design or block randomization. Table F8 shows OLS regression estimates of the relationship between the selected covariates and the outcome of interest in each study.

Table F8: OLS estimates for selected variables in simulation studies

	Study 1	Study 2	Study 3	Study 4	Study 5	Study 6
(Intercept)	0.364*	4.459*	0.128*	0.139*	0.549*	−0.359*
	(0.092)	(0.202)	(0.018)	(0.010)	(0.195)	(0.052)
Z	0.076*	0.407*	−0.012	−0.079*	0.080	0.023
	(0.032)	(0.116)	(0.015)	(0.012)	(0.081)	(0.045)
Y0	0.010	0.117*	−0.016		−0.064	0.001*
	(0.011)	(0.042)	(0.010)		(0.062)	(0.001)
X1	−0.009	−0.649*	0.006	0.000	0.140	0.027
	(0.047)	(0.131)	(0.015)	(0.000)	(0.083)	(0.046)
X2	0.014	−0.127	0.004	−0.001	−0.007*	0.325*
	(0.033)	(0.150)	(0.015)	(0.001)	(0.003)	(0.045)
X3	−0.001	1.030*	−0.010	0.001	−0.103*	
	(0.001)	(0.160)	(0.015)	(0.005)	(0.032)	
X4	0.077	−0.583*	−0.013	0.000	−0.034	
	(0.051)	(0.117)	(0.015)	(0.000)	(0.088)	
X5	−0.005			0.000*	0.150	
	(0.034)			(0.000)	(0.117)	
X6	0.041				−0.192	
	(0.035)				(0.112)	
X7	0.080*				0.331	
	(0.034)				(0.238)	
Num.Obs.	946	2784	561	2712	275	1175
R2 Adj.	0.009	0.056	−0.002	0.022	0.080	0.045

* $p < 0.05$

HC2 standard errors in parentheses. See the pre-registration materials for additional details.

We use these coefficients to determine the predictiveness in column in Table 5 of the article. We classify the blocking variables from Study 2 as "high" because 3 out of 5 variable have large coefficients relative to the treatment effect. We classify the blocking variables from studies 4 and 6 as "moderate" since only one variable in each have large coefficients. Otherwise, we classify the blocking variables from studies as having "low" predictiveness.

We also note two additional deviations from the pre-registration:

- For Study 2, we originally selected six categorical blocking covariates: `race`, `partyID`, `age`, `female`, `education`, and `income`. As we prepared the simulations, we realized that making blocks by intersecting these categories would lead to several groups with too few observations. Therefore, we ultimately decided to only block on binary indicator of `race` (white vs. others), `partyID` as originally coded (democrat, independent, republican), and `female` as originally coded.
- For Study 3, the replication data uses numbers to represent groups for the four categorical variables we selected. For three out of four variables, we could not map the numbers to the categories discussed in the article's supplemental materials, so we created blocks arbitrarily:
  - `religion` ranges from 1 to 11, skipping 3 and 10. We created two groups, the first one for 1 and 2, and everything else in the second, this creates groups that are roughly of equal size.
  - We assumed `education` was coded incrementally, so we created two groups divided by the median value.
  - The majority of observations in `marital` as classified as "2", which we assume corresponds to being married, while the other categories refer to different forms of being non-married. We created two groups accordingly.

F.2 Pre-registered results

A consequence of inadvertently selecting variables with low predictiveness, using them as poxies of pre-treatment outcome measures and/or blocking variables did not lead to precision gains. Therefore, we decided it would benefit the reader if we used simulation to vary the predictiveness of pre-treatment variables in the simulations reported in the main text.

For the sake of transparency, we present simulation results as pre-registered:

- Figure F5 shows the variance ratio of the alternative design against the standard design for different designs under different degrees of sample loss. This figure is analogous to that presented in Figure 2 of the main text. Values below one indicate that the alternative design exhibits lower variance in the distribution of estimated ATEs, which would suggests the corresponding alternative design improves precision even after sample loss. This is only the case for Study 2, especially when incorporating all the blocking covariates. In the other studies, since the variables we chose have a weak relationship with the outcome, we see no improvements in precision. In turn, this implies that sample loss leads alternative designs to perform worse than the standard design.
- We pre-registered reporting the Kolmogorov-Smirnov test statistic as a non-parametric alternative to the F-test. Figure F6 reports these results for the original simulation. This test would help in detecting nonparametric shifts in the distribution of estimated ATEs. We find that the KS statistic is distinguishable from zero both when the alternative design performs better or worse than the standard design. We uncover no hidden patterns that may have gone unnoticed by resorting to a non-parametric test statistic
- We also pre-registered reporting statistical power descriptively to aid in the interpretation of F-tests. Figure F7 shows these results. Except for Study 2, we find that sample loss harms the performance of alternative designs, but since the standard design already has low power, the difference is not considerable. In Study 2, we could have observed alternative designs improving over the standard design if it wasn't for the fact that the standard design already maximizes power, meaning there is not much room to improve.

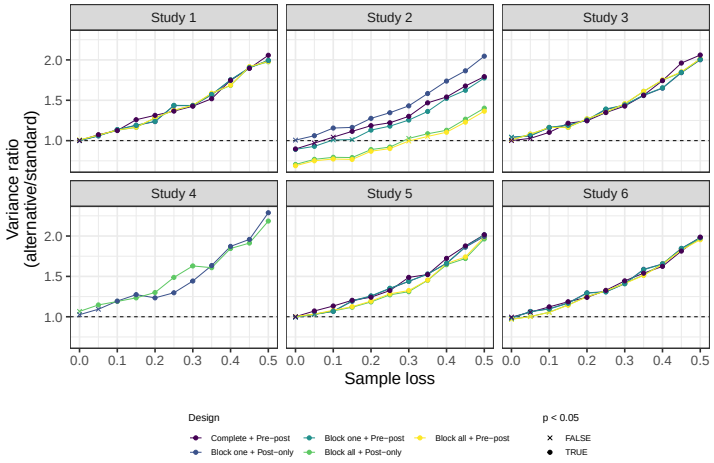


Figure F5: F -test pre-registered simulations results

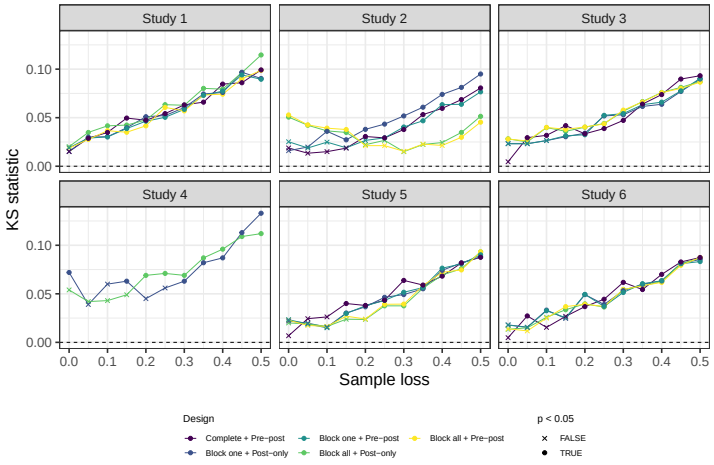


Figure F6: KS-test pre-registered simulations results

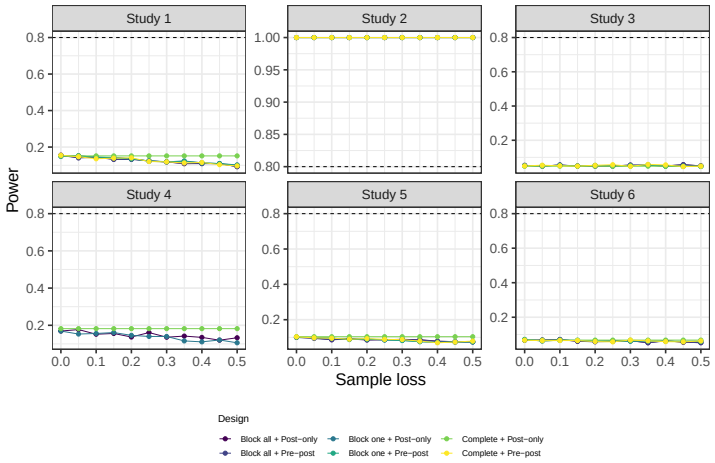


Figure F7: Pre-registered simulations results for statistical power

G Additional Simulation to Assess Full Sample Loss and Pre-treatment Covariate Correlation Space

In this Appendix, we build on our simulation studies in Section 6 of the article. We provide an in-depth simulation study using data from another published experiment to explore the full space of potential sample loss and pre-treatment covariate correlations. Specifically, we use data from Anspach and Carlson (2020). In this experiment, Anspach and Carlson assess the extent to which social media users believe misinformation in the comments of a news post that contains factually-correct information. The authors find evidence that misinformed comments can cause people to retain the incorrect information, despite the correct information being available. Using the authors' replication archive, we imagine how an applied researcher might navigate block randomized and pre-post design decisions in this context.[3]

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

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

To assess the impacts of alternative design choices in the presence of sample loss, we take the authors' reported treatment effects as truth, simulate potential outcomes from their esimated model, and assess how different hypothetical designs would perform in terms of precision. Specifically, the authors model trust in the poll referenced in the news article or preview controlling for several pre-treatment covariates. We use these reported coefficients when simulating potential outcomes. In other words, we *assume model four in their reported results is the true state of the world*, and simulate potential outcomes according to it:

$$\begin{aligned} Y_{i,t2} = 3.28 - .18 * \text{Preview}_i - .57 * \text{Liberal Comment}_i - .62 * \text{Conservative Comment}_i \\ +.02 * \text{Need for Cognition}_i - .01 * \text{Need for Affect}_i - .05 * \text{Knowledge}_i + \\ .01 * \text{Age}_i + .01 * \text{White}_i + .02 * \text{Education}_i + .01 * \text{Income}_i - .45 * \text{Party}_i + u_i, \end{aligned} \tag{2}$$

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

For the following simulation, we focus on the average treatment effect of the preview only condition compared to the full article

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

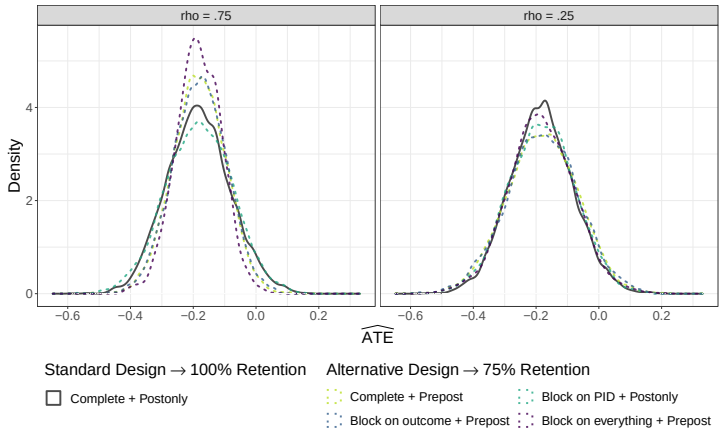


Figure G8: Alternative designs improve precision even with sample attenuation

condition. We focus on this $\widehat{ATE}$ because the authors find it does not reach conventional levels of statistical significance, although it is close ($p = 0.064$). With potential outcomes simulated according to Equation 2, we then simulate the different design decisions outlined in Table 1. In all, we assess the following five designs:

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

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

Table G9: Power of standard design and alternatives with sample loss

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

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

The left plot in Figure G8 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 G8, we can see how these results are contingent upon the pre-treatment information being highly predictive of the outcome. In this plot, the simulated pre-treatment measure of the outcome is only weakly correlated ($\rho = .25$) with the post-treatment measure. Now, the precision gains from incorporating this information cannot overcome the precision losses from losing sample size.

To consider the information in Figure G8 differently, consider the power of the standard design and its alternatives reported in Table G9. 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.

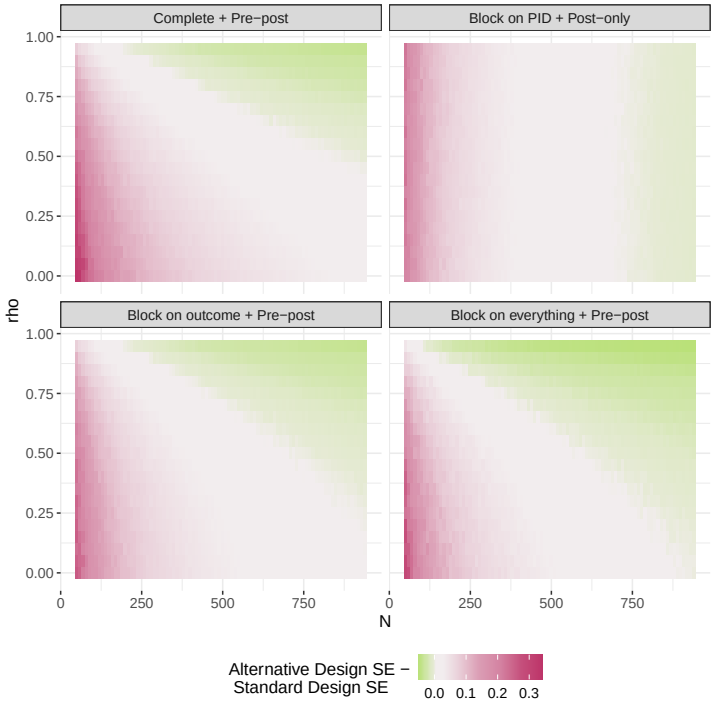


Figure G9: Regions of N and ρ where alternative designs can improve precision

Finally, the key feature of this application is that we extend the simulation in Figure G9 to consider *all* combinations of potential sample loss and predictive quality of pre-treatment covariates. For each of the four alternative designs we consider, we simulate $SE(\widehat{ATE})$ across the nearly full range of sample retention (N $= [.05N, .95N]$) and the full range of how well the pre-treatment measure of the outcome predicts the outcome ($\rho = [0, .95]$). Figure G9 visualizes results as a heatmap comparing how alternative designs compare to $SE(\widehat{ATE}_{\text{Standard}})$, where we consider the standard design to be one with 100% sample retention and the pre-treatment and post-treatment measure of the outcome correlated at $\rho = .75$. Green regions represents combinations of N and ρ where the alternative design does better than the standard design, white regions represent where the alternative design does no better or worse, and red regions represent the alternative design does worse than the standard design.

In the bottom left panel, we see blocking on all pre-treatment information available and using a pre-treatment and post-treatment outcome measure correlated at $\rho = .5$, we find that Anspach and Carlson could have precision *gains* even if their sample size dropped to only 680 participants (29% sample loss). In general, in line with the literature's advice, we find that the more pre-treatment information a researcher utilizes in their design, the more precision gains they can see (bottom right panel). If the pre-treatment information is not highly predictive of the outcome, the gains in precision it brings are not likely to be worth the costs in precision from sample attenuation. We see this in the top right plot where only some sample loss can occur before the decision to block on partisan identity actually harms precision. Finally, we show that if the pre-treatment and post-treatment measure of the outcome are highly correlated, incorporating this information is likely to improve precision, even if it costs the overall sample size.

H Non-Random Sample Loss

In this Appendix, we expand on the article's discussion of non-random sample loss. To facilitate exposition, the application and simulation in the main article text assume that sample loss happens at random. This implies that choosing to invest in block randomization or a pre-post design depends only on the tradeoff between statistical precision and sample retention. However, if sample loss were to systematically affect some units over others, then one should worry about the possibility of alternative designs inducing bias in the estimation of average treatment effects. The results in Appendix Sections E.5 and E.6 of this appendix suggest that this is not a major concern in the studies we selected to replicate.

To extend our treatment of sample loss to a broader setting, we first discuss the circumstances under which sample loss may induce bias. Then, we present simulations that may assist researchers in determining whether bias from sample loss is of concern in their experiment.

H.1 When Does Sample Loss Introduce Bias?

From a bias perspective, one can think of sample loss as equivalent to attrition or missing outcomes. This implies that sample loss induces bias only when it correlates with potential outcomes, meaning that the pattern of missing outcomes correlates with how units respond to treatment (Lo, Renshon, and Bassan-Nygate 2023). This is a problem for navigating the balance between precision and retention when the bias would appear under an alternative design, but not under the standard design.

As an illustration, consider the case of a researcher pondering whether to measure one additional pre-treatment variable for precision

gains via blocking or pre-post design. The concern is that measuring this variable may lead some units to drop from the study in a pattern that induces bias in estimates.

To identify the circumstances in which differential attrition may lead to bias, consider an experimental design with pre-treatment Z, covariate X, and outcome Y. Figure H10 presents directed acyclic graphs that illustrate three possible patterns.

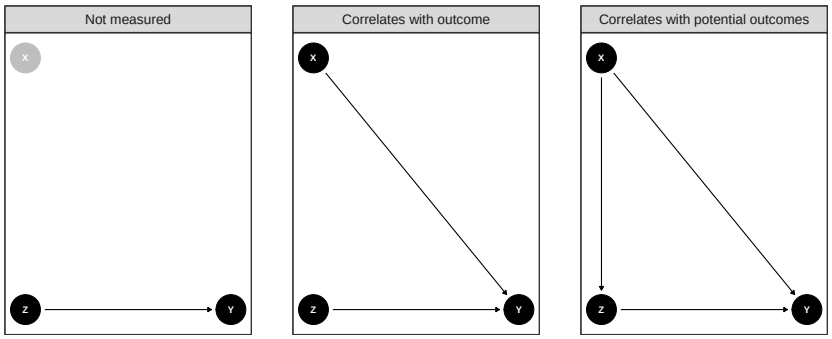


Figure H10: Potential sample loss patterns

We have 3 possible scenarios, arranged from left to right: (1) Covariate X is not measured, (2) Covariate X is measured and only correlates with the outcome, and (3) Covariate X is measured and correlates with potential outcomes (this is represented with an arrow going to both treatment Z and outcome Y)

In the first case, the covariate is never measured so no alternative design is implemented, sample loss cannot happen, and therefore there is no bias. In the second case, the covariate *is* measured and differential attrition happens. But because the relationship between the covariate and the outcome is independent from the relationship between outcome and covariate, differential attrition does does not induce bias. This situation resembles a pre-post design, since pre-treatment measures of the outcome should correlate with the outcome, but not with potential outcomes. In the third case, the covariate is measured and differential attrition happens. However, X now correlates with how units *react* to the treatment, which is equivalent to saying that it correlates with potential outcomes. This means that differential attrition induced by measuring covariate X may induce bias in the estimation of the average treatment effect. This may happen with block randomization, since the idea is to deliberately choose blocking covariates that one expects to correlate with potential outcomes.

The implication from Figure H10 is that sample loss from implementing an alternative designs is only likely under block randomization or a similar design that relies on measuring pre-treatment covariates that can correlate with potential outcomes (e.g. covariate adjustment). This form of bias is not a problem under implicit sample loss scenarios, since in that case the costs are internalized by the researcher, missing outcomes are hypothetical, and treatment assignment happens after measuring outcomes.

Correlated attrition can be a problem for studies in which explicit attrition is a concern. For example, the measurement of pre-treatment variables in a single wave survey may alert respondents to the topic of the study, which may lead them to engage with experimental vignettes differently and, in turn, to attrit at different rates across treatment and control conditions. Sheagley and Clifford (2023) find no evidence that measuring moderators, usually good candidates for blocking, alter treatment effects, which implies that sample loss, if it exists, does not induce bias. For two-wave survey an field experiments, sample loss can correlate with potential outcomes if treatments are assigned before measuring pre-treatment covariates. However, this can be solved by randomizing between waves. Rerandomization is also option when assigning treatments before the first wave is necessary (Li, Ding, and Rubin 2018).

H.2 Incorporating Bias from Sample Loss in Simulations

While we argue that bias from sample loss in our context is rare, we present simulations to illustrate how to incorporate concerns about bias while balancing precision and retention. We simulate experiments in a similar fashion to section main text. However, we only consider the post-only standard and post-only block randomized experiment as alternatives. We ignore pre-post designs here since one would not expect pre-treatment outcomes to affect potential outcomes, whereas one usually chooses to block randomizes on covariates that are highly predictive of potential outcomes. However, the exercise here should also apply to pre-post designs.

For simplicity, consider an experiment with only two blocks, and assume that sample loss happens in only one of the two blocks. This would be the most straightforward way in which measuring pre-treatment covariates for block randomization may induce bias through sample loss. Furthermore, we allow the true treatment effect τ to vary across blocks to show how the problem only emerges when sample loss correlates with potential outcomes. We consider two scenarios for $\tau = (0.2, 0.2)$ and $\tau = (0.3, 0.1)$. Since units are distributed evenly across blocks, the true average treatment effect is the same for the entire sample, but sample loss will only correlate with potential outcomes in the second example. For each combination of parameters and research designs, we simulate 1,000 independent experiments.

Figure H11 shows the power and bias of the four possible combinations of research designs and treatment effect patterns over a range of sample loss rates. The left panel shows how power changes as a function of sample loss. The block randomized experiment is generally more precise than the standard experiment, this is because the underlying blocking covariate is always correlated to potential outcomes.

Because potential outcomes do not correlate with sample loss under the first pattern of treatment effects $\tau = (0.2, 0.2)$, the block randomized experiment still retains high power at high degrees of sample loss, this is because the one block that does not drop

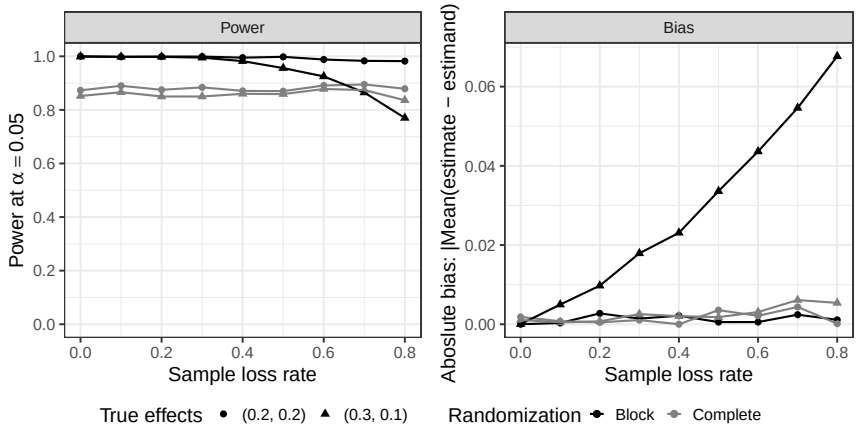


Figure H11: Statistical power and bias for simulated experiments along sample loss rate

observations still retains a sufficiently large sample size. Power only suffers under the pattern of treatment effects that correlates with sample loss, $\tau = (0.3, 0.1)$. This is because more and more units from the first block are being dropped, which according to the right-hand side panel leads to increasing absolute bias in the estimation of the overall ATE. In this stylized simulation, the improvement in statistical precision is justifiable even at the cost of sample loss and bias. For example, under the second pattern of treatment effects, losing about 40% of the observations in the first block leads to a bias slightly above of 0.02 standard deviations in the standard normal outcome.

While the decision of how much bias to tolerate will depend on the specifics of each application, the simulation exercise here suggests that, in general, one should not worry about correlated sample loss when choosing alternative designs any more than one should worry about potential bias from attrition in general.

I Simulating Alternative Designs and Sample Loss Without Existing Data

Throughout the article, we evaluate the performance of alternative designs under sample loss by replicating previously published studies and using data from published experiments as a template to conduct simulations. In this section, we present simulations that evaluate performance under sample loss that do not require access to existing data. This may help researchers to articulate their initial thoughts about the merits of collecting additional pre-treatment variables. The functions to conduct this type of simulation are available in the `simprecision` R package.

I.1 Setting

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

We assign a binary treatment to half of the sample via complete randomization. For the designs that include block randomization, treatment assignment is completely randomized within blocks with the same proportion of treated units in each block. The potential outcome under control $Y_{i,t2}(0)$ a standard normal distribution and correlates with $Y_{i,t1}$ with $\rho = 0.8$. The potential outcome under treatment is $Y_{i,t1}(1) = Y_{i,t2}(0) + \tau Z_i + X_i$ where $\tau = 0.2$ is the true ATE and $Z_i = \{0, 1\}$ denotes treatment assignment. We choose the value of N and τ so that the standard design has middling power, meaning there is room to improve by considering alternative designs.

Since potential outcomes correlate both with the pre-treatment outcome and covariate, we expect any combination of pre-post measurement and block randomization to improve in terms of power in the absence of sample loss. To illustrate this balancing act, we simulate different experiments with varying sample loss rate ranging from 0 to 0.8. We assume two things about sample loss. First, we assume that sample loss happens at random. In some contexts, as in the case of experimental attrition, sample loss may correlate with potential outcomes, which can induce bias to ATE estimation. Appendix E discusses this issue in more detail. In short, we show that this form of correlated sample loss further complicates the balance of precision and retention by adding bias to the mix, yet we argue that one should not worry about this issue any more than what one should worry about experimental attrition in general. Second, we assume the standard design never suffers from sample loss. Moreover, sample loss is the same regardless of the alternative research design under consideration. In some contexts, this assumption may not be realistic, since the cost of measuring covariates and outcomes may differ. However, this assumption is sufficient to convey sample loss as a function of the proportion of observations the researcher expects to lose.

Our second scenario follows the same setting, but keeps the sample loss fixed at 0.25 while varying N from 100 to 5,000 observations. This illustrates the case when entertaining an alternative design implies changing the target sample size. For example, adding a baseline survey may force the researcher to study 500 units instead of 1,000, while still losing a quarter of the sample between waves, leading to an effective sample of $500 - 125 = 375$. For each combination of parameters, we simulate 1,000 experiments and estimate the ATE using the corresponding estimator for each design: The difference in means for the standard design, the difference-in-differences for the

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

I.2 Results

Figure I12 shows the distribution of statistical power for our two simulation scenarios. The left panel shows the statistical power of the standard and alternative experimental designs. Since the standard design (complete randomization + post only) does not suffer from sample loss, its power is constant around 0.6 along the horizontal axis. This serves as a benchmark to evaluate whether it is worth investing in alternative research designs. Holding everything else constant, implementing either a block randomization post-only (block + post only) design or a pre-post design under complete randomization (complete + pre-post) improves upon the standard design as long as the researcher expects to lose less than 30% of the sample. Combining both pre-post outcome measurement and block randomization (block + pre-post) improves precision even further, and its power exceeds the standard design as long as the researcher expects to lose less that 60% of the sample.

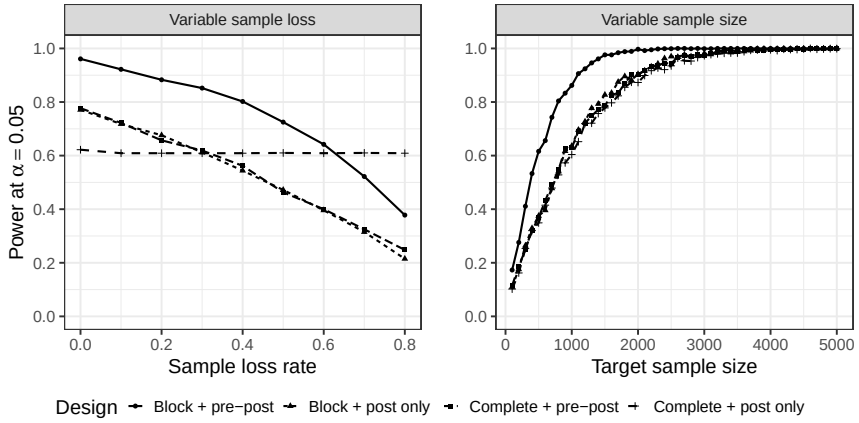


Figure I12: Statistical power for simulated experiments along sample loss rate and target sample size

The translation of these findings to concrete practices depends on the nature of the application. For example, if investing in alternative designs implies collecting a sample that is half as small as the sample one would collect under the standard design, as in the case of an experiment that requires baseline and endline surveys, then only the block + pre-post design leads to an improvement in terms of precision. Another way to interpret the results is to compare power horizontally. For example, implementing only one of block randomization or pre-post measurement without any sample loss has roughly the same power as an experiment that loses 40% of the sample by using both. Yet another way to interpret the results would be to interpret sample loss in terms of how much smaller of a sample a researcher can afford by investing in a new design. For example, one can afford to collect 60% fewer observations by implementing a block + pre-post design and still achieve comparable levels of precision.

The right panel of Figure I12 shows power for the same alternative designs, but fixing sample loss at 25% and varying the target sample size, which means that the effective sample size is smaller for the alternative designs. For example, a complete + pre-post design with a target sample size of $1,000$ ends up collecting information for $1,000 - 250 = 750$ observations. At this rate of sample loss, the increase in precision from implementing either a complete + pre-post or block + post only is offset by the reduction in the effective sample size. Given these parameters, only the block + pre-post combination leads to a net increase in power. If we focus on the conventional target of 80% power, an experiment that loses 25% of the observations by deviating from the standard design can achieve this with around 800 observations under a block + pre-post design, and with around 1500 observations otherwise, including the standard design.

These findings depend exclusively on the parameters we chose for our stylized simulations, but they convey three types of conclusions that researchers can draw by entertaining the choice of alternative designs at the pre-analysis stage. First, if the goal is to entertain explicit sample loss emerging from the marginal cost of measuring one additional variable, then the application is more likely to exist in the domain of the vertical comparisons in the left panel, and the question is whether one would be willing to sacrifice a small to moderate decrease in sample size to increase precision. Second, if the goal is to address implicit sample loss from being forced to collect a smaller sample, then the horizontal comparisons in the left panel are more relevant. The question is then how much of the sample loss associated with conducting an additional wave in data collection is tolerable in terms of preserving the target statistical power. Finally, if the goal of entertaining alternative designs is to minimize data collection costs while preserving statistical power while accounting for both kinds of sample loss, then the comparisons in the right panel can be used to determine what would be the minimal target sample to collect under alternative research designs.

J Flowchart

Ultimately, the decision to collect additional pre-treatment information and implement an alternative design needs to be considered on a case-by-case basis. We intend for this article to make the practical and statistical components of this decision clearer so researchers are better equipped to consider these design choices. We broadly summarize our advice and findings in a flowchart in J13.

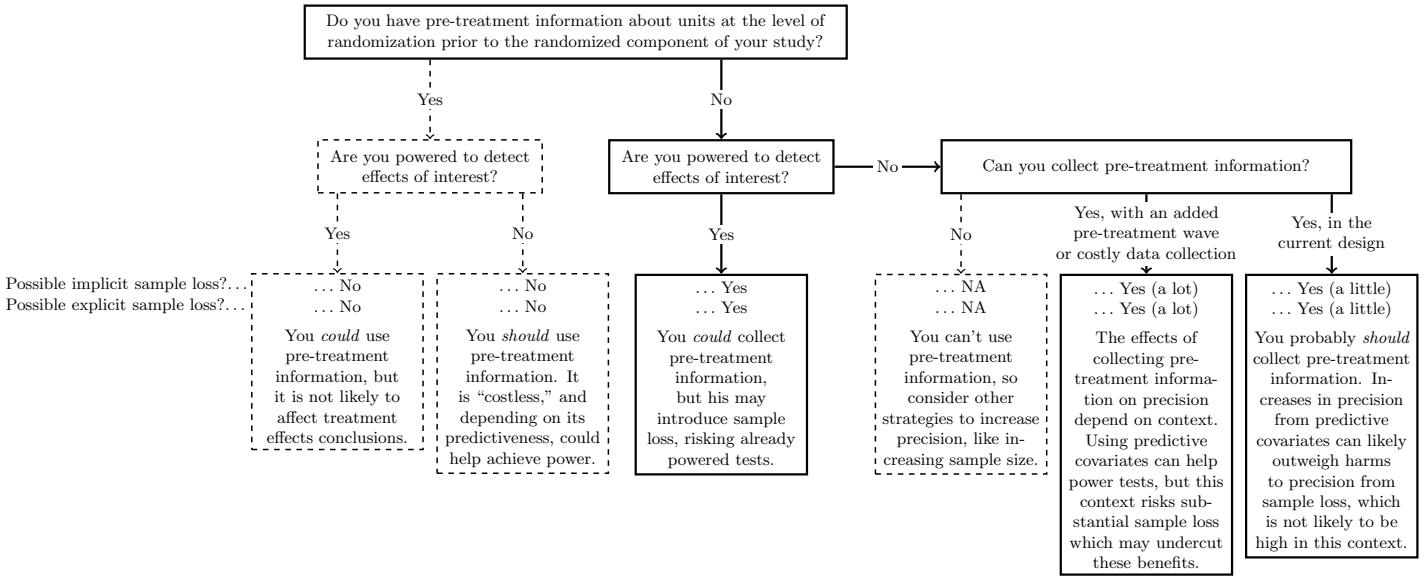
The first question a researcher must answer is whether they already have pre-treatment information about the experimental units at their disposal. For example, if the experimental units are classrooms, the researcher may already know prior test scores, demographics, and more about these classes. In this case, neither explicit nor implicit are a concern, and the researcher can use the pre-treatment information they already have to increase statistical precision.

The discussion in this article focuses on the right path answering "No" to the question of whether pre-treatment information is available. If so, the researcher should next consider if their treatment effect estimation is powered. If the researcher is confident they have sufficient power, they *could* use pre-treatment information to increase their precision. However, the researcher needs to carefully consider any possible implicit or explicit sample loss that may result from collecting additional pre-treatment information. If the predictiveness of pre-treatment covariates is not strong, and significant sample loss is incurred, the alternative design could result in precision *losses.* Our evidence provides a cautionary tale. The Dietrich and Hayes (2023) replication had a small beginning sample size and a quasi pre-treatment measure of the outcome with low correlation (r=0.28), and our simulation showed block randomization lost any precision gains after losing 10% of the sample. Similarly, Study 6 in our simulations had a small sample, and only pre-treatment variables that are moderately or highly correlated with the outcome showed precision gains under sample loss. If already powered, alternative designs that incur sample loss pose a risk.

We next consider what a researcher might do if they answer "No", they are not powered to detect effects of interest. In this case, the researcher must consider any avenue available to them prior to fielding their experiment to increase precision. As we discuss in this article, block randomization and pre-post designs are strongly encouraged in the literature with promises to improve precision, but the researcher needs pre-treatment information about experimental units to implement these designs. The next box in the flowchart considers the feasibility of collecting such information. We outline three possibilities.

First, a researcher may not be able to collect pre-treatment information. For example, a researcher may not have access to their sample prior to randomization. For a design like Munger (2017), the experimental intervention was randomly assigned in real time when a user posted a racist Tweet. In this case, if a researcher lacks precision in their estimates, they ought to consider other strategies to increase precision we mention in the article.

Figure J13: First Steps in Considering Whether to Implement an Alternative Design to Increase Precision



Note: The left path indicating "Yes" uses dashed arrows and boxes. This denotes the literature's existing advice regarding the benefits of incorporating pre-treatment information into a design. Right path with bolded boxes indicate summarized advice from this article.

Second, if the researcher can collect pre-treatment information, but it requires an additional pre-treatment wave or other costly activities, like paying for or hiring a research assistant to collect pre-treatment information about the geographic clusters in one's field experiment, the researcher must consider the possibility of both implicit and explicit sample loss. For example, D. Broockman and Kalla (2016) implement a separate pre-treatment survey in their field experiment studying the effects door-to-door canvassing can have on decreasing transphobia. It is possible that they could have afforded more people in the experimental phase of the study, but implicitly sacrificed sample size in order to collect pre-treatment information (D. E. Broockman, Kalla, and Sekhon 2017). Moreover, explicit sample loss is a major concern with implementing a pre-treatment wave. Units will likely drop between waves, possibly outweighing precision gains of incorporating pre-treatment information into the design. In sum, a researcher in this context should carefully consider implementing an alternative design. It may not be worth collecting this information because sample loss might be substantial and not outweigh the design choices' benefits. However, in some cases using pre-treatment information could make the difference between being powered or underpowered, even if sample loss occurs. For example, our evidence from the Tappin and Hewitt (2023) replication shows the sizable precision gains of an alternative design requiring a pre-treatment wave, even under significant sample loss.

Third, we consider the scenario in which the researcher can collect pre-treatment information without implementing a new pre-

treatment wave. Instead, they can collect pre-treatment information using the structure of their current design. For example, online survey experiments that randomize participants to conditions within the survey can easily add additional pre-treatment measures into the design. As our replication evidence shows, this setup is unlikely to feature explicit sample loss under alternative designs. The negative consequences of implicit sample loss can also be kept to a minimum in this context. Survey time may increase by adding pre-treatment questions, and a researcher may not be able to afford as many units as a result. But if the predictiveness of the covariates is moderate or strong, our evidence repeatedly shows that precision gains are likely to withstand minor sample loss that might stem from asking additional questions.

References

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

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

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

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

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

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.

Broockman, David, and Joshua Kalla. 2016. "Durably Reducing Transphobia: A Field Experiment on Door-to-Door Canvassing." *Science* 352 (6282): 220–24.

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.

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.

Freedman, David A. 2008. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics* 40 (2): 180–93. https://doi.org/10.1016/j.aam.2006.12.003.

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.

Gibbons, Charles E., Juan Carlos Suárez Serrato, and Michael B. Urbancic. 2018. "Broken or Fixed Effects?" *Journal of Econometric Methods* 8 (1). https://doi.org/10.1515/jem-2017-0002.

Humphreys, Macartan. 2009. "Bounds on Least Squares Estimates of Causal Effects in the Presence of Heterogeneous Assignment Probabilities." *Manuscript, Columbia University.*

Li, Xinran, Peng Ding, and Donald B. Rubin. 2018. "Asymptotic Theory of Rerandomization in Treatment–Control Experiments." *Proceedings of the National Academy of Sciences* 115 (37): 9157–62. https://doi.org/10.1073/pnas.1808191115.

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

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

Munger, Kevin. 2017. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment." *Political Behavior* 39: 629–49.

Sheagley, Geoffrey, and Scott Clifford. 2023. "No Evidence That Measuring Moderators Alters Treatment Effects." *American Journal of Political Science*, July. https://doi.org/10.1111/ajps.12814.

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.