Balancing Precision and Retention in Experimental Design

Appendix

May 10, 2023

Contents

A. Flowchart for Applied Researchers

An experimenter's decision to collect additional pre-treatment information needs to be carefully considered on a case-by-case basis. We intend for this article to make the practical and statistical components of this decision clear so a researcher is better equipped to consider this design choice in their studies. To aid in the first steps of this process, in this section we include a flowchart to summarize the advice and findings in the article that researchers can consult.

The central question this article seeks to help researchers answer is whether they should collect more pre-treatment information. How might this affect the competing components of precision? Will collecting this information be beneficial to use for a blocking or pre-post design? Or, will collecting this information require or lead to a smaller sample size, thus harming the researcher's ability to detect treatment effects?

Figure A1 presents a flowchart to help balance these competing concerns. The first question posed is whether the researcher already has pre-treatment information that they can incorporate in their design. Note that 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. The right path indicating "No" is where this article contends with the unclear state of advice in the literature, and unpacks how to consider the benefits of alternative designs when they might simultaneously prompt sample loss.

Existing Advice in the Literature

First, consider the left path indicating "Yes," the researcher has pre-treatment information. For example, a cluster randomized design might randomly assign some classrooms to an intervention and some to a placebo control. This researcher may already know prior test scores, demographics, and more about these classes. Or, university labs may have an existing participant pool with pre-collected demographic data.

Continuing down this branch, the next pertinent question the researcher should ask themselves is whether or not they are powered to detect their effects of interest. In this article, and in most political science experiments, researchers are foremost interested in average treatment effects. If the researcher is powered, they could use pre-treatment information to increase their precision. Indeed, when researchers only have one shot at estimating treatment effects, we advise they take every reasonable measure to increase precision in order to detect true treatment effects. However, incorporating this information into the design may not affect final conclusions about treatment effects of interest if the researcher is confident their tests are powered.

If a researcher's tests are not powered, we highly recommend they use the pre-treatment information at their disposal. Using pre-treatment information (i.e., block randomization and pre-post outcome measurement) could stand to greatly increase precision, and depending on the predictiveness of the covariates, the precision gains could be substantial. Indeed, in this context, there is no implicit sample loss to worry about – the pre-treatment information is free so the researcher does not need to sacrifice sample size ex ante to pay to gather it. Moreover, there is no explicit sample loss to worry about – the pre-treatment information is already collected so the researcher does not need to worry about attrition during the study as a result of collecting it.

Beyond Existing Advice

Now, consider right path of the flowchart, indicating that the researcher does not have pre-treatment information. For example, a survey experiment conducted using participants on Amazon's Mechanical Turk would lack individual-level information about the participants before fielding the study.

As before, continuing down this branch, the researcher should then consider if their tests are 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 pursing the collection of additional pre-treatment information.

We next consider what a researcher might do if they answer “No”, they are not powered to detect effects of interest. It is critical the researcher consider any avenue available to them prior to fielding their experiment to increase power. As we discuss in this article, block randomization and pre-post designs are strongly encouraged in the literature with promises to improve precision. Critically, 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 common 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) implements, 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 discuss in the article. The simplest strategy is to increase sample size as much as possible.

Finally, the flowchart considers the central question in this article—what a researcher should consider when they feasibly could collect additional pre-treatment information and implement alternative designs, like block randomization and pre-post measurement, to increase precision in their estimates.

If it is feasibile, but it requires an additional pre-treatment wave, 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 to use when estimating treatment effects (D. E. Broockman, Kalla, and Sekhon 2017). Moreover, explicit sample loss is a major concern. Units will likely drop off between a pre-treatment wave and experimental wave of a study. Another study may consider these competing components of precision and decide the sample loss is not worth implementing these alternative designs that require pre-treatment information. In sum, a researcher in this context (underpowered and questioning whether to collect pre-treatment information) will have to carefully consider this decision. It not be worth collecting this information because sample loss might be substantial and not outweigh the design choices' benefits. However, using pre-treatment information could stand to greatly increase precision. Depending on the predictiveness of the covariates, it could make the difference between being powered or underpowered, even if sample loss occurs.

Finally, we consider the last branch of the flowchart. In this instance, a researcher does not need to implement 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. This context is not likely to feature large implicit or explicit sample loss. Survey time may increase by adding pre-treatment questions, and a researcher may not be able to afford as many units as a result (implicit loss), but if the predictiveness of the covariates is high, precision gains are likely to withstand minor sample loss. Moreover, it is unlikely that many units drop due to survey fatigue from a few added pre-treatment questions. In this setting, it is fairly safe to assume that the precision gains from incorporating pre-treatment information, and designs like block randomization and pre-post measurement, are going to outweigh harms to precision from sample loss.

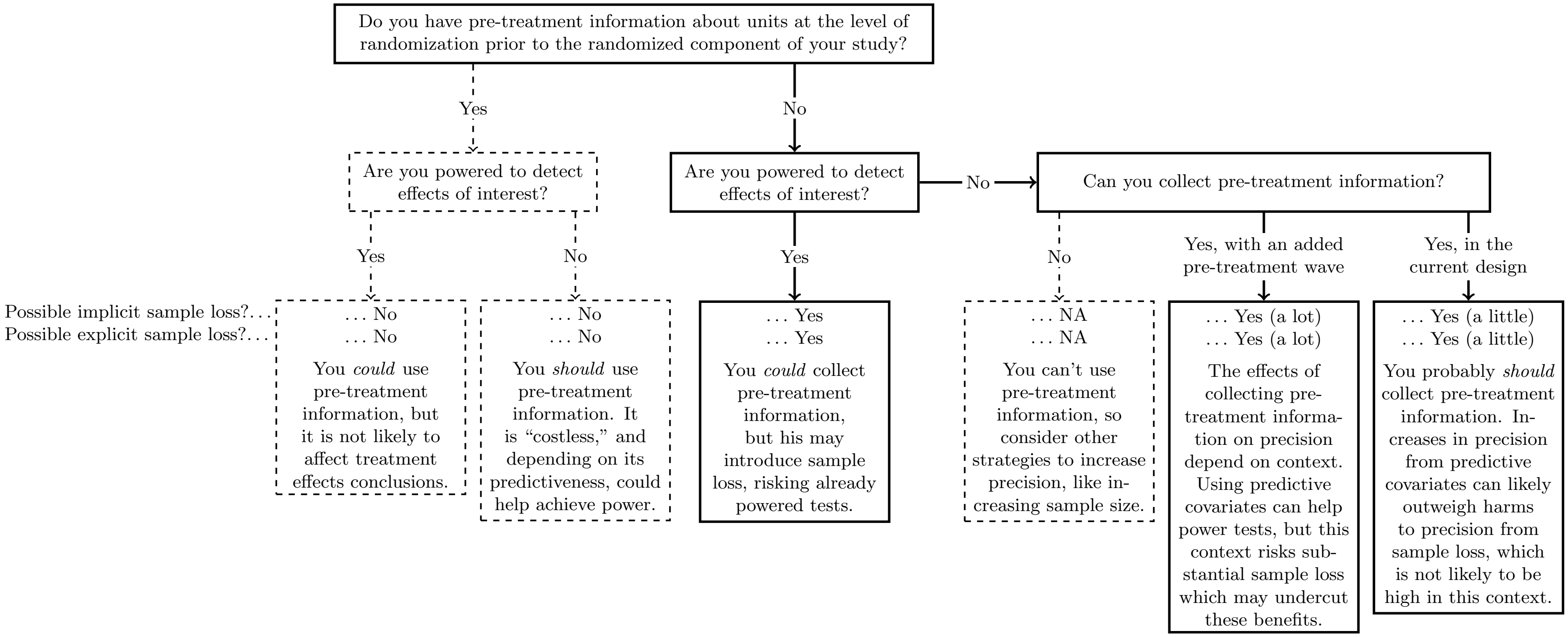
Do you have pre-treatment information about units at the level of randomization prior to the randomized component of your study?
Yes
No
Are you powered to detect effects of interest?
Are you powered to detect effects of interest?
No
Can you collect pre-treatment information?
Yes
No
Yes
No
Yes, with an added pre-treatment wave
Yes, in the current design
Possible implicit sample loss?. . .
Possible explicit sample loss?. . .
. . . No
. . . No
You *could* use pre-treatment information, but it is not likely to affect treatment effects conclusions.
. . . No
. . . No
You *should* use pre-treatment information. It is "costless," and depending on its predictiveness, could help achieve power.
. . . Yes
. . . Yes
You *could* collect pre-treatment information, but his may introduce sample loss, risking already powered tests.
. . . NA
. . . NA
You can't use pre-treatment information, so consider other strategies to increase precision, like increasing sample size.
. . . Yes (a lot)
. . . Yes (a lot)
The effects of collecting pre-treatment information on precision depend on context. Using predictive covariates can help power tests, but this context risks substantial sample loss which may undercut these benefits.
. . . Yes (a little)
. . . Yes (a little)
You probably *should* collect pre-treatment information. Increases in precision from predictive covariates can likely outweigh harms to precision from sample loss, which is not likely to be high in this context.

Figure A1: First Steps in Considering How Alternative Designs Balance Precision and Retention

B. 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 B1: 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.

[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.

Using the example in Table B1, 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, $Var(Y_i(0))_2 = .25$, $Var(Y_i(1))_2 = .25$, $Cov(Y_i(0)_2, Y_i(1))_2 = .25$, $N = 2$, 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.

C. 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.

D. Estimators for Pre-Post Designs

The main text 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

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i \tag{2}$$

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, as in our application to Anspach and Carlson (2020), 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 using the conventional OLS standard errors hurts precision.

E. 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.

In this context, sample loss is equivalent to attrition or missing outcomes, which can induce bias in two ways. First, attrition may turn representative samples into non-representative samples, which challenges external validity. Our paper is not concerned with external validity since we focus primarily on strategies to improve statistical precision to enhance internal validity. We direct readers to Findley, Kikuta, and Denly (2021), Egami and Hartman (2022), and Lo, Renshon, and Bassan-Nygate (2023) for recent treatments on the subject. The general advice there applies to the research designs we discuss as well.

Second, sample loss may induce bias when it correlates with potential outcomes, meaning that the pattern of missing outcomes may correlate 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.

This form of correlated attrition would happen in our setting when collecting pre-treatment data on outcomes or blocking covariates leads units to abandon the study after randomization with higher frequency in some experimental conditions over others. Which is not a problem for implicit sample loss 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.

This turns the problem from a balance of precision and retention into a bias-precision-retention tradeoff, which complicates the choice of optimal experimental design even further. One can be in a position where an alternative design suggests considerable improvements in statistical precision while inducing non-negligible bias. This implies getting estimates that are more consistent yet further away from the true ATE.

To illustrate how to incorporate bias concerns into the balance, 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.

The setup is identical to that described in section 5, except that now we 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.

Figure E1 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.

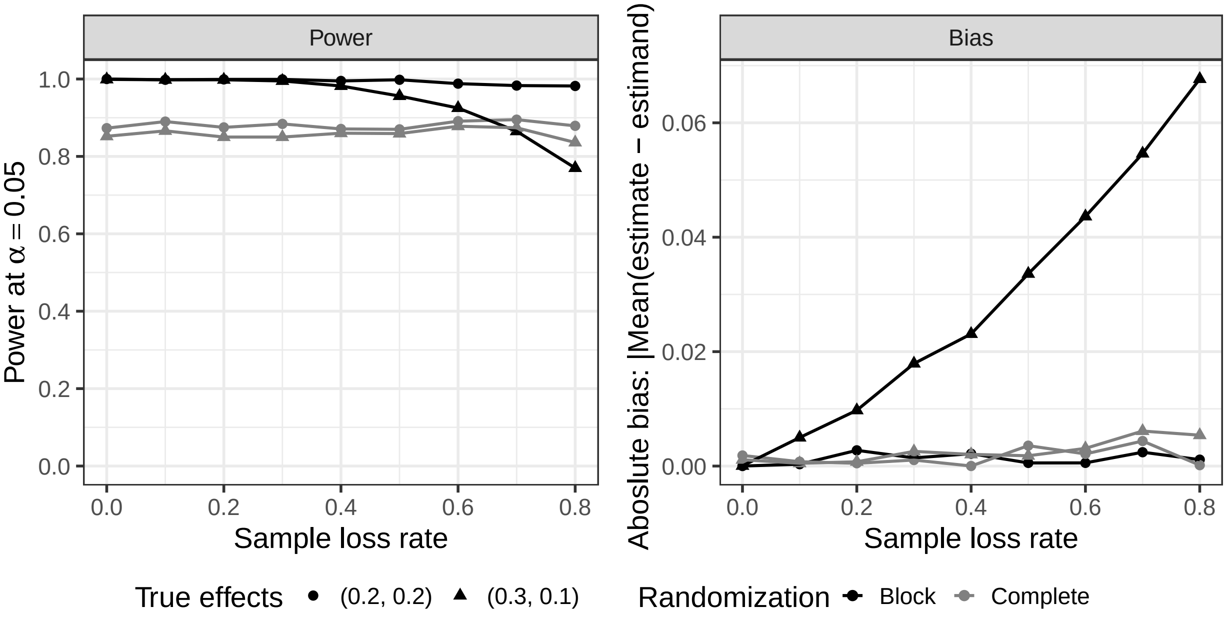

Figure E1: Statistical power and bias for simulated experiments along sample loss rate
Note: Each point along the horizontal axis is based on 1,000 simulated experiments.

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 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.

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. 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.

Egami, Naoki, and Erin Hartman. 2022. "Elements of External Validity: Framework, Design, and Analysis." *American Political Science Review*, October, 1–19. https://doi.org/10.1017/s0003055422000880.

Findley, Michael G., Kyosuke Kikuta, and Michael Denly. 2021. "External Validity." *Annual Review of Political Science* 24 (1): 365–93. https://doi.org/10.1146/annurev-polisci-041719-102556.

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.*

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.