Manufacturing Missingness: A Hidden Cost of Survey Experiments

Tara Slough*

June 26, 2025

Abstract

Survey experiments are often used to measure the effects of interventions embedded within surveys. Current practice strongly favors between-subjects experiments (in which some subjects are shown one treatment while others are shown another treatment) over non-experimental within-subjects designs (in which participants see both treatments, one after another) because between-subjects designs invoke weaker assumptions for unbiased estimation of aggregate treatment effects. However, this practice neglects the major cost of between-subjects designs: the loss of information about individual responses to interventions. I argue that researchers should carefully evaluate the bias-variance tradeoff when designing intervention-oriented surveys. To do so, I propose the first non-heuristic decision rules to navigate this tradeoff in light of researchers' substantive goals. These rules can be used prospectively or via proposed hybrid survey designs that measure the quantities relevant to the decision rules. Holding fixed the number of intervention-oriented surveys, researchers would be well served to experiment less frequently.

*Associate Professor, New York Unviersity. tara.slough@nyu.edu. I am grateful to Scott Clifford and Alex Coppock for sharing additional documentation and replication files from their articles. I thank Neal Beck, Alex Coppock, Sandy Gordon, Jiawei Fu, Dorothy Kronick, Justin Melnick, Kevin Munger, Marcus Prior, Mike Tomz, and Carolina Torreblanca for helpful comments on an early version of this manuscript. Thanks also to seminar audiences at Stanford University, the Princeton Center for the Study of Democratic Politics "Advances in Measurement in Survey Experiments" workshop, and NYU WINNING.

Suppose that a political scientist or pollster wanted to know if calling an "inheritance tax" a "death tax" makes the public more favorable to eliminating the estate tax. The typical method for answering this question is a survey experiment where a subset of respondents is randomly assigned to the death tax language and the remaining respondents see the more neutral wording. The researcher or pollster would then compare aggregate (e.g., average) support for elimination of the estate tax between each treatment group. In current practice, few scholars would reach for a design where everyone sees both wordings and we look at within-respondent differences in opinion. In this paper, I argue that the latter design is often superior to between-subjects survey experiments and propose tools for determining which design is preferable in a given application.

Over the past three decades, political scientists have rapidly adopted survey experiments as a workhorse method for answering questions in all empirical subfields. Following Gaines, Kuklinski, and Quirk (2006), by survey experiment, I refer to experimental research designs in which: (1) randomly assigned treatments or interventions are delivered on a survey instrument or by a survey enumerator and (2) outcomes are measured on the same survey instrument. Since the first political science survey experiments in the early 1990s (Sniderman and Piazza, 1993; Sniderman, 2012), survey experiments have become the modal quantitative methodology across 174 political science journals (Grossman, Dinneen, and Torreblanca, 2025). Within the 20 journals with the highest impact factors, nearly 8% of quantitative empirical articles were survey experiments in 2023 (Grossman, Dinneen, and Torreblanca, 2025), and in two general-interest journals, survey experiments constituted 15-20% of *all* articles by 2023 (Briggs et al., 2025). Moreover, the high prevalence of survey experiments in pre-analysis plans and conference applications (Figure 1) suggests that the growth of survey experiments in journals is not an artifact of selection during the publication process.

Amid this rapid growth in the use of survey experiments, a large literature has emerged to provide methodological guidance to survey experiment practitioners (e.g., Gaines, Kuklinski, and Quirk, 2006; Hainmueller, Hopkins, and Yamamoto, 2014; Mummolo and Peterson, 2019; Clifford, Sheagley, and Piston, 2021; Offer-Westort, Coppock, and Green, 2021; Brutger et al., 2022;

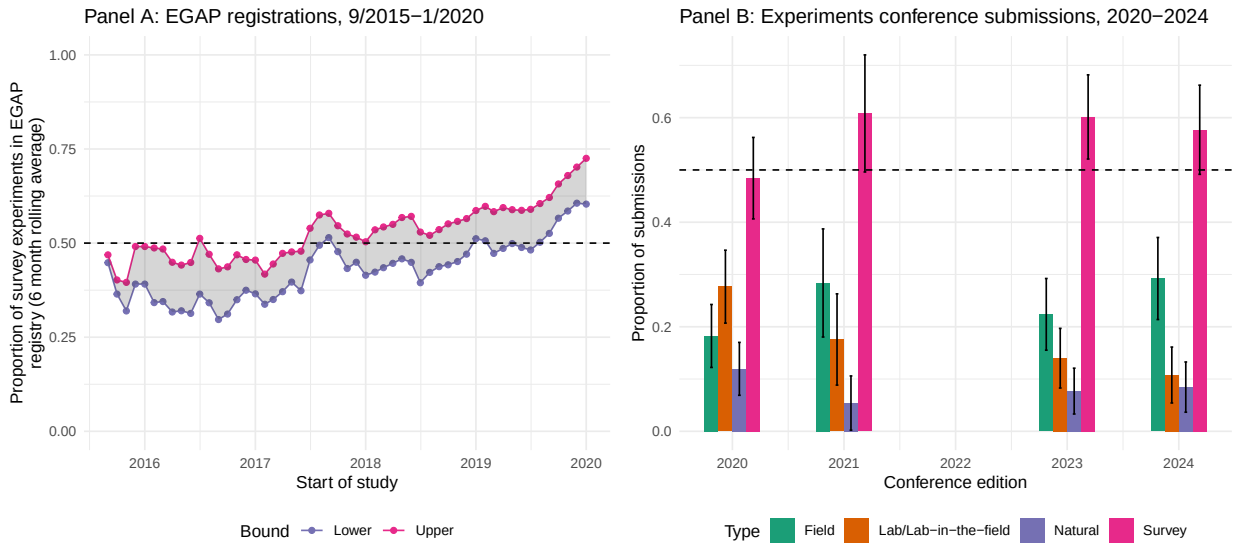

Figure 1: Two estimates of the prevalence of survey experiments among experiments political science. Panel A includes pre-registration information for 1,198 studies beginning between September 2015 and January 2020. Panel B includes 506 conference applications for conferences held in 2020-2024.

Kertzer, 2022). But with few exceptions, namely Blair, Coppock, and Moor (2020), this literature presumes that it is optimal to administer treatments experimentally. I argue instead that many studies that are currently implemented as survey experiments should not be *experiments* at all.

I compare between-subjects survey experiments to within-subjects designs in which individual subjects view multiple treatment conditions in the same survey. As in the estate tax example, in the (between-subjects) survey experiment, subjects are assigned to a question using either the "inheritance tax" ($Z = 0$) or "death tax" ($Z = 1$) wording. Both treatment groups are asked for their support for the estate tax to measure outcomes $Y(Z)$. This design is depicted in the left panel of Figure 2. In contrast, in the within-subjects design, subjects might first be asked about their support for the estate tax with the "inheritance tax" wording ($Z = 0$). Denote this outcome as $Y_1(0)$, where the subscript 1 corresponds to the order of the treatment. Subjects would then be asked about their support for the estate tax given the "death tax" wording ($Z = 1$), in order to measure the outcome $Y_2(1)$, where the subscript 2 corresponds to the ordering of the treatment. Of course, with this example, the ordering of treatments could be reversed or even randomized, as depicted in the three survey flows in the right panel of Figure 2.

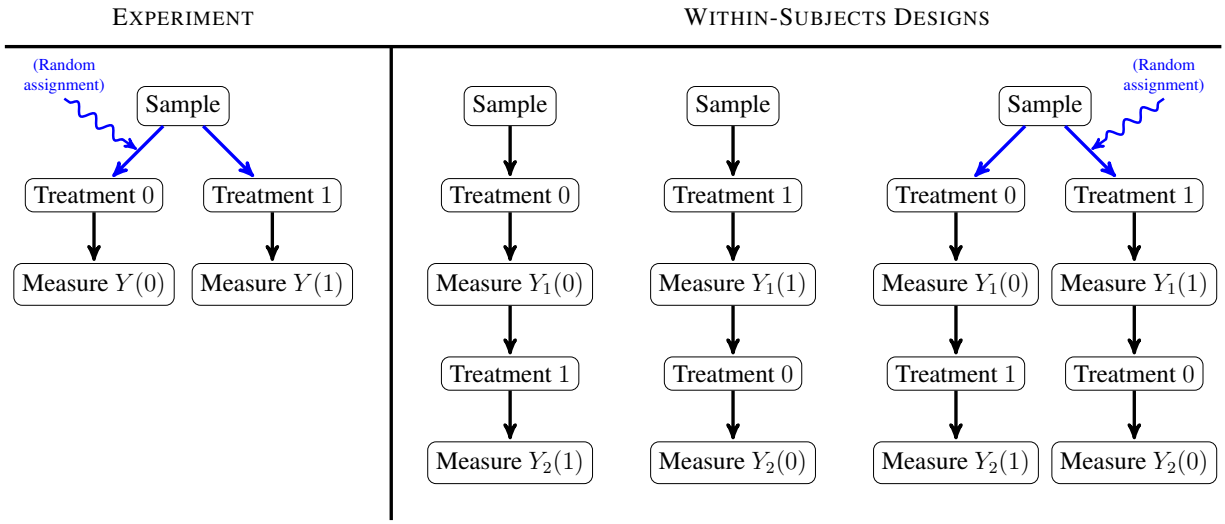

Figure 2: Simplified survey flow for experimental and within-subjects designs. Not all interventions can be logically implemented with a randomized within-subject design, so the right panel depicts three possible within-subject designs for a binary intervention.

Ultimately, the researcher wants to measure differences between $Y(1)$—support for the estate tax with the "death tax" wording—and $Y(0)$, support for the estate tax with the "inheritance tax" wording. The experimental design permits only comparisons between subjects in different treatment conditions. In contrast, the within-subjects designs allow for measurement of differences between $Y_t(0)$ and $Y_{\neg t}(1)$ for each individual subject.[1] By considering within-subjects designs as a class of research designs that makes fundamentally non-experimental comparisons, I broaden substantially the class and applicability of these designs relative to experimental within-subjects designs (e.g., crossover designs). This substantially reduces questions about their scope for applicability due to the difficulty or impossibility of "undoing" or removing treatments that subjects have already read (Clifford, Sheagley, and Piston, 2021).

My argument is premised on an influential analogy between causal inference and missing data problems (Ding and Li, 2018). Causal inference problems can be seen as problems of missing counterfactual potential outcomes. In the experimental design in Figure 2, we are missing measures of $Y(1)$ for all subjects assigned to treatment 0 and vice versa. The within-subjects designs

[1]The within-subjects design with randomized ordering depicted in Figure 2 admits several potential experimental comparisons, a point to which I will return in Section 5.

do not solve the problem of missing counterfactual potential outcomes: we do not observe $Y_1(1)$ for subjects assigned to treatment 0 at $t = 1$ (etc.). But it may well be the case that $Y_2(1)$ provides at least some information about the unobserved $Y_1(1)$. Ultimately, when either type of design is feasible, choosing between an experimental and the within-subjects designs boils down to our tolerance for missing data (in the experimental design) or assumptions about how potential outcomes measured at different t may be related to each other (in the within-subjects design).

Survey (and lab) experiments are unique because both experimental and within-subjects designs are often feasible. When both designs are feasible, comparison of these designs should be viewed as a bias-variance tradeoff. The (between-subjects) experimental design relies on less data and treatment effect estimates therefore generally have higher variance. For example, when treatment effects are independent of (control) potential outcomes, a two-arm experimental design with half of the sample in each arm requires a minimum of *double* the sample size of the within-subjects design to achieve the same power for the average treatment effect (ATE).[2] In many plausible settings, the experimental design requires substantially more than double the sample size to achieve equivalent power. I refer to this feature of experimental designs as "manufacturing missingness."

On the other hand, as is widely appreciated, non-experimental within-subjects designs risk the introduction of bias. There are multiple mechanisms that could generate this bias. Priming and differential demand effects are widely discussed (Charness, Gneezy, and Kuhn, 2012; Clifford, Sheagley, and Piston, 2021). Less appreciated are the possibility of respondent fatigue and the implications of outcomes that are not stationary. While we do not know the degree to which these biases present in a given application, existing explorations should give some pause to the idea that these phenomena are always present (e.g., Klein et al., 2014; Mummolo and Peterson, 2019; Bansak et al., 2019).

I propose the first non-heuristic solution to the bias-variance tradeoff that draws upon researchers' substantive inferential goals. When researchers seek to test a directional prediction, my recommended decision rule indicates that they should weigh bias sufficient to "flip" the sign

[2]This best case for the experimental design is achieved only when half of subjects are assigned to each arm and the distribution of individual treatment effects has infinite variance.

of the estimate against the precision gains of the within-subjects design. When researchers seek instead to precisely estimate a treatment effect (or another quantity of interest), I suggest an alternate decision rule, in which researchers choose the design that minimizes mean-squared error of the estimator. While the logic is similar, these two decision rules can select different designs. For this reason, clearly articulating the substantive goals of an intervention-oriented survey is essential for optimizing an empirical research design.

Researchers can use my proposed decision rules with a hypothetical data generating process or with simulated data to determine whether or not to experiment. This application is in line with the power calculations that researchers regularly conduct; indeed, many of the relevant parameters in the decision rules are parameters of standard power calculations! Alternatively, I discuss two hybrid survey designs that nest experimental and within-subjects designs, allowing researchers to estimate the parameters of the relevant decision rule to inform the choice of primary research design. Relative to a (pure) within-subjects design or a (pure) experimental design, the hybrid approach has lower statistical power. Simulations show, however, that the hybrid approach allows researchers to select the design consistent with their chosen decision rule (on average) without making strong *ex-ante* assumptions about how subjects will respond to a given treatment or sequence of treatments.

Through an original survey employing a hybrid design and re-analysis of data from an existing within-subjects (crossover) design by Clifford, Sheagley, and Piston (2021), I show that hybrid designs are widely applicable and straightforward to employ. The results from these applications find strong performance of the within-subjects designs over standard between-subjects experiments, though this performance may vary across applications.

This paper contributes to discussions of survey/survey experimental design. It draws upon considerations of increasing efficiency in survey experiments through repeated measures that are surveyed by Clifford, Sheagley, and Piston (2021). However, in contrast to this work, it introduces the possibility that experiments may not be the optimal way to answer substantive questions of interest. This finding echoes the treatments of list experiments versus direct questions on sensitive

subjects by Blair, Coppock, and Moor (2020). By formalizing the tradeoffs between experimental and non-experimental design and proposing formal decision rules, this paper improves upon the heuristic decision rules forwarded by Charness, Gneezy, and Kuhn (2012) and List (2025).

More broadly, this paper explores a tension between two central tenants of the credibility revolution: the emphasis on research design (over assumptions) and a preference for unbiased estimators (Slough and Tyson, 2024). The preference for unbiasedness is typically presented as a lexicographic preference (e.g., Esterling, Brady, and Schwitzgebel, 2024). I show, however, that by being more explicit about how we trade off bias and variance, a design-based perspective sometimes favors observational over experimental research designs for learning about causal effects that measure substantive phenomena of interest.

1 Causal Inference and Missing Data: An Analogy

This paper starts from existing observation of a close link between problems of causal inference and problems of missing data. In the potential outcomes framework, a causal effect is defined as a difference in potential outcomes evaluated at different values or levels of a treatment. Measuring causal effects is difficult due to the fundamental problem of causal inference. The fundamental problem of causal inference holds that it is impossible to observe the more than one potential outcome (at multiple levels of a treatment) for each unit (Holland, 1986). Ding and Li (2018: p. 214) make an analogy between the fundamental problem of causal inference and missing data problems writing: "because for each unit at most one of the potential outcomes is observed and the rest are missing, causal inference is inherently a missing data problem."

Table 1 illustrates the link between the fundamental problem of causal inference and missing data in the case of a study with a binary treatment, $Z_i \in \{0, 1\}$. For each unit (indexed by i), the treatment takes a value normalized to 0 or 1 (e.g., "control" or "treatment"). Whereas potential outcomes are defined for each unit under each treatment (by assumption), we can only observe one potential outcome y_i for each unit. The remaining cells are marked with a question mark. This question marks are analogous to missing observations. As a result, we cannot observe the

individual causal effect at the unit level. However, if we can randomize the assignment of Z_i, thereby conducting an experiment, we can measure an aggregate causal effect by aggregating over multiple units.

Unit (i)	Z_i	$Y_i(1)$	$Y_i(0)$
A	1	y_A	?
B	0	?	y_B
C	0	?	y_C
D	1	y_D	?

Table 1: Assignment to a binary treatment is represented by $Z_i \in \{0, 1\}$. A measured outcome variable is denoted y_i. The question marks indicate potential outcomes that are unobserved under the realized treatment assignment.

It is worthwhile to consider a different study design, however. Suppose that we could measure potential outcomes for each unit by providing each unit with both versions of the treatment. Obviously, we cannot not expose units to each treatment condition simultaneously. But, suppose that it were possible to provide subjects with one treatment condition, measure the associated potential outcome, and then provide subjects with the other treatment condition and then measure that potential outcome. Here, we must recognize that potential outcomes could change between the first outcome measurement and the second. Moreover such changes could be endogenous or exogenous to either treatment condition. In other words, even in our hypothetical design, there is no claim to have surmounted the fundamental problem of causal inference.

Table 2 depicts the schedule of potential outcomes for a within-subjects design for a study with a binary treatment, analogous to the study in Table 1. Here, potential outcomes $Y_{it}(Z_{it})$ are indexed by unit (i) and the order in which a treatment is administered/outcome is measured (t). The indicator O_i indexes the order in which the treatments are administered. $O_i = 1$ indicates that the "control" condition is measured first and the "treatment" condition is measured second, and $O_i = 0$ indicates the reverse. There are now two distinct potential outcomes for each treatment condition given the ordering of the treatments.[3]

[3]Note that it may be the case that $Y_{i2}(Z_{i2})$ varies as a function of the treatment at time 1, Z_{i1}. In

Unit (i)	O_i	$Y_{i1}(1)$	$Y_{i2}(1)$	$Y_{i1}(0)$	$Y_{i2}(0)$
A	1	?	y_{a2}	y_{a1}	?
B	1	?	y_{b2}	y_{b1}	?
C	0	y_{c1}	?	?	y_{c2}
D	0	y_{d1}	?	?	y_{d2}

Table 2: Alternate design. The order in which the treatments are administered is given by $O_i \in \{0, 1\}$, where $O_i = 1$ means that $Y_i(0)$ is measured first. A measured outcome variable is denoted y_{it}, where t indexes the order of the measurement. The question marks indicate potential outcomes that are unobserved under this design.

It is clear from Table 2 that the within-subjects design does not solve the fundamental problem of causal inference: there are still question marks in the table. Specifically, we measure $Y_{it}(0)$ and $Y_{it}(1)$ at different t's. The important questions becomes: *for all units (i) and any treatment (Z), can we learn about $Y_{i1}(Z)$ from $Y_{i2}(Z)$?* In the extreme case, suppose that we assumed that potential outcomes did not vary in t. This assumption can be formalized as follows:

Assumption 1. *For all units $i \in N$ and treatments $z \in Z$, $Y_{it}(Z) = Y_i(Z)$ for all $t \in T$.*

Under Assumption 1, the design in Table 2 permits the identification of individual treatment effects (ITEs) as follows:

$$ITE_i = Y_{it}(1) - Y_{i\neg t}(0) = \begin{cases} y_{i2} - y_{i1} & \text{if } O_i = 1 \\ y_{i1} - y_{i2} & \text{if } O_i = 0. \end{cases} \tag{1}$$

In the context of our analogy to missing data, the assumption that potential outcomes do not vary in t allows us to impute missing potential outcomes using the realization of potential outcomes at other t. In this instance, we clearly obtain more information—ITEs—from doing the within-subjects study. If we were willing to invoke Assumption 1, we would strictly prefer a within-subjects design to the between-subjects experiment on grounds that the within-subjects design contains more information.

this case, the potential outcome could be written $Y_{i2}(Z_{i2}, Z_{i1})$. I discuss this possibility in Section 3.

The assumption that potential outcomes do not vary in t is strong. Yet, the converse—that we cannot learn anything about the unobserved potential outcome $Y_{it}(z)$ from the observed potential outcome $Y_{i\neg t}(z)$—is arguably equally as strong. From a missing data perspective, the latter assumption holds that we gain nothing from using observed potential outcome $Y_{i\neg t}(z)$ to impute unobserved potential outcome $Y_{it}(z)$. It is thus useful to examine the properties of these designs to understand when each design may be desirable.

1.1 Practical consideration: the plausibility of a within-subjects design

Two features of a study are important for establishing the plausibility of a within-subjects design. First, all treatments of interest must be manipulable. This requirement is satisfied in all experiments, whether field, survey, or lab. Not all observational studies have this property, though manipulability (by some actor other than the researcher) is important in many observational studies targeting causal effects. This feature is therefore not particularly restrictive.

Second, a within-subjects design requires that a subject be exposed to more than one treatment condition in sequence. Outcomes must further be measured in response to each treatment. This distinguishes many survey and lab experiments from other types of experimental designs. Consider the case of a field experiment testing a large-scale development intervention that runs for several years. A within-subjects design that tests multiple treatment conditions in each community is likely to be extremely costly. Moreover, it would raise some concerns about how unobserved potential outcomes evolve over the extended period of the interventions. In contrast, many survey experimental manipulations can be implemented in relatively rapid succession on a survey instrument. This is analogous to evaluating multiple treatment conditions in a single lab session. Despite the close analogues, while within-subjects designs have been embraced in some lab experiments (Charness, Gneezy, and Kuhn, 2012), they remain rare in survey experiments. Indeed, Clifford, Sheagley, and Piston (2021) identify only one within-subjects intervention-oriented survey among 67 published survey experiments in five leading political science journals.

An original hand-coding of the survey experiments assembled by Clifford, Sheagley, and Piston (2021) identifies five broad classes of treatments used in existing survey experiments (in de-

Treatment type	Description	Proportion	Within-Subjects design considerations
Vignette	Varies characteristics of a *hypothetical* person/group of people, event, or phenomenon.	53.8%	Often amenable to repeated vignettes (in any order) for each subject. However, some vignettes also reveal (non-hypothetical) information that cannot be withdrawn.
Information	Varies what information is revealed about a person/group of people, event, or phenomenon.	40.0%	Cannot "remove" information after it is revealed, limiting the application of designs that randomize the ordering of treatments.
Question wording	Varies wording of a question on the survey.	15.4%	Often amenable to repeating treatment conditions for each subject in any order. However some question wordings also reveal information that cannot be withdrawn.
Incentives	Varies material payouts or strategic considerations for respondents	7.8%	Often amenable to repeating treatment conditions for each subject in any order.
Priming/emotional induction	Varies stimulus to induce different emotions, mental states, or cognitive processes	6.1%	Repeated induction relies on how transitory induced emotions/states/cognitive processes are.

Table 3: Original classification of treatments in 67 experiments assembled by Clifford, Sheagley, and Piston (2021). Note that some treatments fall into multiple categories, so the proportions sum to more than 100%. See Appendix A3.

creasing order of frequency): vignettes, information, question wording, incentives, and priming or emotional induction. I describe these classes of articles and their frequency in Table 3. Some classes of treatments logically restrict the order in which treatments can be administered. For example, if treatments that reveal real information about a person, event, or phenomenon, it is infeasible to remove that information to study an individuals' responses absent the information.[4] As a consequence, it is not necessarily possible to randomly assign the order of treatments (or present treatments in any order) in many of these experiments. This property has been cited as a barrier to the applicability of within-subjects *experiments* in political science (Clifford, Sheagley, and Piston, 2021). However, it is not necessarily an impediment to the use of a within-subject study design.

[4]We do not know how long primed emotions last in studies that induce/prime an emotion or state of mind (Gillies and Dozois, 2021). Our ability to induce multiple emotions in a single survey instrument depends on how long such induced emotional states last.

2 The Cost of Experimentation: Efficiency Loss

We will first consider the efficiency gains from a within-subjects design relative to the experiment. To isolate these gains—the benefit of the within-subjects design—we will assume that Assumption 1 holds. This means that we can express potential outcomes without reference to an order or time subscript. We will first focus on the minimum sample size required to achieve power of $1 - \beta \in (0, 1)$. Existing rules of thumb set $\beta = 0.2$ (for power of 0.8) or $\beta = 0.1$ (for power of 0.9).

Suppose that we want to know the effect of a single binary treatment, $Z_i \in \{0, 1\}$, on an outcome $Y_i(Z) \in \mathbb{R}$. Further, within the experiment, we will assume complete randomization of treatment with equal probability of assignment to each arm. We will denote individual treatment effects by τ_i, which implies that:

$$Y_i(1) = Y_i(0) + \tau_i. \tag{2}$$

It is straightforward to see that the ATE is given by $E[Y_i(1) - Y_i(0)] = E[\tau_i]$. Consider the following estimators of the ATE within each design:

$$\text{Within-subjects design:} \qquad \overline{Y_i(1) - Y_i(0)} \tag{3}$$

$$\text{Experiment:} \qquad \overline{Y_i(1|Z_i = 1)} - \overline{Y_i(0|Z_i = 0)} \tag{4}$$

In the experiment, we only observe the potential outcome corresponding to each treatment condition. In the within-subjects design, we observe $Y_i(Z)$ for both treatments. The within-subjects design estimator is unbiased if Assumption 1 holds and the experimental difference-in-means estimator is unbiased if assignment of Z_i is ignorable (which is supported by the randomization of treatment).[5]

Given the potential outcomes defined above, we can straightforwardly derive standard errors

[5] Additional assumptions of excludability and SUTVA follow from the way in which potential outcomes are defined above.

of each ATE estimator, as follows:

$$\text{Within-subjects design:} \quad se_w^{ATE} = \sqrt{\frac{\text{Var}[\tau_i]}{N}}$$

$$\text{Experiment:} \quad se_e^{ATE} = \sqrt{\frac{\text{Var}[Y_i(0)] + \text{Var}[\tau_i] + 2\text{Cov}[Y_i(0), \tau_i]}{N/2} + \frac{\text{Var}[Y_i(0)]}{N/2}}$$

$$= \sqrt{\frac{2\text{Var}[\tau_i] + 4\left(\text{Var}[Y_i(0)] + \text{Cov}[Y_i(0), \tau_i]\right)}{N}}$$

Both expressions follow from (2).[6] Recall that variance is (weakly) positive but covariance can be negative or positive. Therefore, by bounding covariance at its theoretical minimum (when it is negative) and comparing these standard errors, it is straightforward to derive the following result:

Proposition 1. *Suppose that Assumption 1 holds. For any data generating process in which* $\text{Var}[Y_i(0)] > 0$ *and/or* $\text{Var}[Y_i(1)] > 0$, $se_e^{ATE} > se_w^{ATE}$.
(All proofs in Appendix.)

While Assumption 1 is strong, the other conditions in the proposition—that $\text{Var}[Y_i(0)] > 0$ or $\text{Var}[Y_i(1)] > 0$—are satisfied in effectively all experiments. Further, note that while the expression for se_e^{ATE} above describes experiments in which treatment is assigned with probability $1/2$, the proof of Proposition 1 allows for lopsided allocations of treatment. The loss of efficiency in the experiment comes from two sources that are evident from comparing the standard errors above. First, in the experiment, we only observe each potential outcome for a half of the sample. This is a consequence of "manufacturing missingness." Second, to the extent that potential outcomes are correlated within-subjects we reduce the variance of the ATE estimates by differencing outcomes at the individual level. These relative efficiency gains are comparatively larger when treatment effects are more homogeneous (as $\text{Var}[\tau_i]$ decreases) or when treatment effect heterogeneity is positively correlated with untreated potential outcomes (when $\text{Cov}[Y_i(0), \tau_i] > 0$.)

The gains in efficiency with the within-subjects design have important implications for statistical power. It is worthwhile to ask how many subjects would be necessary to power each design (for

[6] Note that $\text{Var}[Y_i(1)] = \text{Var}[Y_i(0)] + \text{Var}[\tau_i] + 2\text{Cov}[Y_i(0), \tau_i]$.

a given β) given $Y_i(0)$ and τ_i. For power of $1-\beta$, the effect must be $\Phi^{-1}(p = 1-\beta, \mu = 1.96, \sigma^2 = 1)$ z-scores away from zero, where Φ^{-1} denotes the inverse cdf of the normal distribution. Given the standard errors above, we can then solve for N in each design (as follows):

$$\text{Within-subjects design:} \quad \Phi^{-1}(1-\beta, 1.96, 1) = \frac{E[\tau_i]}{\sqrt{\frac{\text{Var}[\tau_i]}{N_w}}}$$

$$\Rightarrow N_w = \frac{\text{Var}[\tau_i]}{(\frac{E[\tau_i]}{\Phi^{-1}(1-\beta,1.96,1)})^2}$$

$$\text{Experiment:} \quad \Phi^{-1}(1-\beta, 1.96, 1) = \frac{E[\tau_i]}{\sqrt{\frac{2\text{Var}[\tau_i]+4(\text{Var}[Y_i(0)]+\text{Cov}[Y_i(0),\tau_i])}{N_e}}}$$

$$\Rightarrow N_e = \frac{2\text{Var}[\tau_i] + 4\left(\text{Var}[Y_i(0)] + \text{Cov}[Y_i(0), \tau_i]\right)}{(\frac{E[\tau_i]}{\Phi^{-1}(1-\beta,1.96,1)})^2}$$

One way to quantify the efficiency gains of the within-subjects design is to evaluate the ratio N_e/N_w. This ratio indicates how many times larger the experimental sample would need to be than the within-subjects design sample to achieve the same statistical power (for the ATE). Proposition 1 shows that $N_e/N_w > 1$. The theoretical upper bound on this ratio is infinity. Table 4 illustrates how this ratio varies as a function of $\text{Var}[\tau_i]$ and $\text{Cov}[Y_i(0), \tau_i]$ while fixing $\text{Var}[Y_i(0)] = 1$.[7] This ratio is highlighted in gray in the table. As $\text{Var}[\tau_i]$ increases, the ratio N_e/N_w decreases because the relationship between control and treated potential outcomes becomes weaker (noisier). When treatments polarize outcomes (when $\text{Cov}[Y_i(0), \tau_i] > 0$), the within-subjects design yields greater efficiency gains over the experimental design.

Another goal of many survey experiments is to examine heterogeneity in ATEs. For example, depending on the substantive domain, researchers may be interested in how a given ATE varies for men and women (Schwarz and Coppock, 2022), Republicans and Democrats (Graham and Svolik, 2020), or ex-ante supporters and opponents of a policy (Guess and Coppock, 2020). It is well known that subgroup conditional ATEs (CATEs) and especially differences in CATEs are less powered than ATEs. Appendix A reports an analogous analysis to the above and shows that

[7] Fixing $\text{Var}[Y_i(0)] = 1$ is equivalent to standardizing the outcome variable by the control group standard deviation, a common practice.

			$E[\tau_i] = 0.1$		$E[\tau_i] = 0.2$		$E[\tau_i] = 0.3$	
$\text{Var}[\tau_i]$	$\text{Cov}[Y_i(0), \tau_i]$	N_e/N_w	Experiment	Within	Experiment	Within	Experiment	Within
0.5	Min.	4.34	1,705	393	427	99	190	44
0.5	0	10	3,925	393	982	99	437	44
0.5	Max.	15.65	6,145	393	1,537	99	683	44
1	Min.	2	1,570	785	393	197	175	88
1	0	6	4,710	785	1,178	197	524	88
1	Max.	10	7,850	785	1,963	197	873	88
2	Min.	1.17	1,840	1,570	460	393	205	174
2	0	4	6,280	1,570	1,572	393	698	174
2	Max.	6.82	10,720	1,570	2,680	393	1,192	174

Table 4: Illustrative minimum sample sizes needed for 80% power. $\text{Var}[Y_i(0)] = 1$, meaning that $E[\tau_i]$ can be interpreted as a standardized effect. All sample sizes are rounded up to the nearest respondent.

the efficiency gains of the within relative to the experimental design are greater than the efficiency gains reported above for the ATE. As such, the analysis of the ATE represents a *lower bound* on the efficiency loss from experimentation.

3 The Cost of the Within Design: (Possible) Bias

The previous section suggests that survey experiments are inferior to a within-subjects design along one important dimension: efficiency. Yet, most experimentalists justify their use of randomized experimental designs through another criterion: the elimination of bias (under the assumptions of excludability and SUTVA). Recall that the above comparison of the efficiency of these two designs invoked Assumption 1, which is sufficient to ensure that the within-subjects design does not generate bias. In practical applications, this of course cannot be guaranteed. When we relax Assumption 1, it is worth thinking about how bias might emerge and present in the within-subjects design.

With respect to the formalization of the within-subjects design in Table 2, consider the following violations of Assumption 1:

$$E[Y_{i1}(1)] \neq E[Y_{i2}(1)] \text{ or } E[Y_{i1}(0)] \neq E[Y_{i2}(0)]. \tag{5}$$

In these cases, we would have two (possibly different) ATEs, as a function of time or question ordering (t), given by $ATE_t = E[Y_{it}(1) - Y_{it}(0)]$, where the expectation is evaluated over units, i. When either condition in (5) obtains, the within-subjects design cannot generate an unbiased estimate of either ATE_t.[8] In order to understand the possibility of bias, then, we must consider the conditions under which either condition in (5) might occur. To this end, I survey known mechanisms that could generate bias in the within-subjects design. I then consider what evidence exists to suggest that these mechanisms are activated in within-subjects designs in general.

3.1 Mechanisms underlying bias

The mechanisms that drive bias in the within-subjects design could be exogenous or endogenous to treatments. First, consider two sources of bias that are endogenous to earlier treatments: priming and experimenter demand effects (Charness, Gneezy, and Kuhn, 2012). **Priming** refers to the possibility that an earlier treatment or outcome measurement question could "activat[e] various mental constructs" (Weingarten et al., 2016: p. 4), thereby affecting subsequent perceptions or behaviors. In the context of a within-subjects design, priming would be a concern if these subsequent perceptions or behaviors were picked up in responses to subsequent treatments. In this context, an earlier treatment (or response to an earlier treatment) would effectively bleed into responses to a later treatment, therefore inducing bias. This is an important concern for within-subjects designs. However, it is important to note that the magnitude and replicability of priming effects is widely debated in both the social psychology (e.g Doyen et al., 2012; Klein et al., 2014; Shanks et al., 2013; Weingarten et al., 2016) and applied political psychology literatures (e.g., Huber and Lapinski, 2006, 2008; Mendelberg, 2008*a*,*b*). These mixed findings suggest that avoiding or discarding within-subjects designs entirely on fears of priming effects is an overreaction.

Demand effects refer to behavior induced by subjects' purposeful efforts to respond or behave in an effort to appease the researcher or support the researcher's hypotheses/goals (Orne, 1962). While demand effects may be present in traditional survey experiments (but see Mummolo and

[8] In the knife edge case in which $E[Y_{i2}(1)] - E[Y_{i1}(1)] = E[Y_{i2}(0)] - E[Y_{i1}(0)]$, $ATE_{t=1} = ATE_{t=2}$. If this were the case, when the order of the treatments is randomly assigned, the within-subjects estimator of the ATE is unbiased.

Peterson, 2019), it is plausible to think that they subjects learn more about researchers' intentions/goals when subjected to a sequence of treatments and repeated questions. Thus, responses that are measured later in a survey may be more strongly subject to demand effects than the initial iteration of the response. Both priming and (magnified) demand effects should be viewed as *endogenous* to treatments in the within-subjects design.

Second, consider the possibility of **respondent fatigue**. The within-subjects design requires that subjects be exposed to more than one treatment condition and respond to questions after each condition. This mechanically increases the length of surveys (holding fixed the number of non-experimental survey items). Krosnick (1991) suggests that fatigue could drive *satisficing* survey-taking behaviors, which generate satisfactory but non-optimal responses. If respondent attention were to decay in this fashion, thereby increasing measurement error as the survey progresses, outcomes measured later may be measured with greater error than those measured earlier in the survey. This could have two (not mutually exclusive) effects: it will increase noise (Berinsky, Margolis, and Sances, 2014) but it could also bias estimates of treatment effects, depending on (i) the pattern of satisficing and (ii) distribution of the outcome variable (see, e.g., Tyler, Grimmer, and Westwood, 2024). Bias driven by (increased) satisficing over the course of a within-subjects design is a consequence of the use of this design. It may or may not be endogenous to the treatments therein.

Finally, it is possible that potential outcomes **vary over time independently of treatment**. This concern has motivated a substantial amount of methodological work on within-subjects designs including cross-over trials in medical research (Jones and Kenward, 2014). When interventions are relatively lengthy or require a washout phase between the administration of different treatment conditions to one subject, we should naturally be concerned about variation in outcomes over time. In the survey experimental context in which treatments can be quickly deployed and outcomes can be efficiently measured, it is less clear how much over time variation we should expect. Empirically, we rarely have over-time measures of $Y_{it}(0)$ in single-shot surveys, so it is hard

to know whether a given potential outcome fluctuates substantially over short periods.[9] If attitudes, for example, did substantially vary over the course of a 10 or 20 minute survey, we may have additional concerns about what the measure is capturing. It is important to note that non-stationary outcomes as a mechanism underlying bias should be viewed as *exogenous* to treatment. It emerges simply because we measure outcomes at different times.

3.2 Probing evidence of bias

The question of whether the within-subjects design introduces bias ultimately will depend on the specific treatments, outcome measures, mode of administration, and respondent characteristics. In contrast to the clear cut efficiency gains of the within-subjects design over an experiment in Section 2, a finding of bias—or lack thereof—in any given experiment is unlikely to be sufficient to confirm or allay these fears in *any* within-subjects administration of treatment.

While considering what we know about biases induced by the within-subjects design, it is important to keep in mind that current practice relies overwhelmingly on experiments over non-experimental within-subjects designs. This suggests a widespread belief that bias may be pervasive in within-subjects designs. This is despite the mixed (at best) evidence for priming and demand effects (Klein et al., 2014; Mummolo and Peterson, 2019) and limited evidence about how satisficing varies in survey length (Bansak et al., 2019).

It is also useful to consider the modal class of within-subjects designs in current practice, conjoint surveys. Users of this design in political science usually follow the identification strategy for the average marginal component effect (AMCE) proposed by Hainmueller, Hopkins, and Yamamoto (2014). This identification strategy *assumes* "stability and no carryover effects" Hainmueller, Hopkins, and Yamamoto (2014: p., 8). This is a variant of Assumption 1 above. Thus, in this case of a within subjects design, political scientists appear to be comfortable with the invocation of an assumption akin to Assumption 1. In Appendix B, I examine whether there is evidence against this assumption by reexamining ordering effects in a set of twelve recent candidate choice

[9] Of course, we could evaluate related measures in a panel survey, but those measurements occur at a lower frequency than in a within-subjects design.

conjoint surveys (Schwarz and Coppock, 2022). In the reanalysis, I find no evidence that the probability that a candidate is selected varies in the order in which a profile is presented, which provides no evidence against the use of this assumption in the conjoint context.

In sum, the within-subjects design has the potential to induce bias in settings in which an experimental administration of a single treatment per respondent does not. Whether the within-subjects design produces bias in any given research design and setting is ultimately an open empirical question. But there are reasons to doubt that these biases emerge—or are large—for all possible contexts, samples, treatments, or outcomes. Replication studies and work measuring these sources of bias suggest that these biases may be less prevalent than current practice—the heavy use of between-subjects experimental designs—would suggest.

4 Trading off Bias and Efficiency

The previous sections suggest that the tradeoff between conventional survey experimental designs and within-subjects designs can be conceptualized as a bias-efficiency tradeoff. This presents an important question: how much bias would a researcher admit in exchange for the efficiency gains of the within-subjects design? While this tradeoff is recognized in existing work, it is typically resolved by choosing one property over the other. For example, Charness, Gneezy, and Kuhn (2012: p. 7) provide a heuristic-based recommendation of (between-subjects) experiments, writing: "we prefer between designs, but recognize the limitations involved...although within-subjects designs look attractive, the researchers need to make the case that the confounds discussed above do not pose a challenge for the results." List (2025: p. 28) similarly advises between-subjects designs unless the threat of bias in the within-subjects design is "low."[10]

Like the guidance offered by Charness, Gneezy, and Kuhn (2012), many credibility-motivated researchers' decisions suggest a lexicographic preference for unbiasedness (Esterling, Brady, and Schwitzgebel, 2024). Despite this widespread commitment, it is rare to find this argument made

[10]Specifically, he argues that the threat of bias is low when the likelihood of violations of "causal transience" (treatment effects do not carry through to subsequent potential outcomes) and "temporal stability" (potential outcomes do not vary over time) are both "low" (List, 2025: p. 28).

explicitly (but see Gerber, Green, and Kaplan, 2004). To see the limits of treating unbiasedness as a lexicographic preference, consider the limiting case of this approach: an unbiased estimator of a treatment effect with infinite variance. In this case, we would not learn anything from a single realized treatment effect estimate! This is an extreme case, but it suggests that we need to take more seriously the tradeoff between bias and variance. To act upon these considerations I develop two decision rules between the experimental and a non-experimental within-subjects design of the form: "If the bias of the non-experimental design is sufficiently small relative to the reduction in variance, use the within-subjects design; if it is not, use the experimental design."

Any decision rule should depend on the substantive goals of researchers. In general, treatment effects—in survey experiments and beyond—are used for two distinct purposes:

1. To **test a directional prediction** about a causal effect. For example, a theory may predict that treatment increases an outcome, such that $Y_i(1) > Y_i(0)$. In the context of a treatment that tries to persuade respondents to support a candidate, this approach tests the prediction that treatment moves respondents toward the candidate of interest.

2. To **obtain a point estimate of a causal effect** to measure a phenomenon, inform a policy decision, or conduct a back-of-the-envelope calculation (etc.) For example, a campaign may be interested using an estimated average treatment effect of a persuasion treatment to inform a cost-benefit calculation when deciding whether to deploy the persuasion treatment in a campaign.

These goals correspond to distinct decision rules, which I formulate. Consider first that our goal is to test a directional prediction. Here, I propose that a researcher comparing the two designs would want to equalize the probability that an estimate takes the "wrong" sign. Thus, if the effect of a given contrast of treatments on an outcome is positive, then we would be willing to admit bias (toward zero) only until the probability that we recover a negatively-signed treatment effect is equivalent under the two designs. Panel A of Figure 3 depicts the comparison of interest by plotting the sampling distributions of the estimators under each design. The experimental design—in

black—generates an unbiased (but noisy) estimate of the treatment effect, τ.[11] We would be indifferent between the experimental (black) and within-subjects designs (dark red) when the density below zero is equivalent (since $\tau > 0$ in the graph). Namely, our goal is to determine how large the bias b_s (s for sign) must be to outweigh the variance reduction. It is straightforward to show that this indifference occurs whenever:

$$\underbrace{\frac{\tau}{se_e}}_{\text{Experiment}} = \underbrace{\frac{\tau + b_s}{se_w}}_{\text{Within-subjects design}}$$

Remark 1 follows from rearranging this expression and provides the conditions under which the within-subjects design is preferred in terms of b_s. Recall that our goal in this setting is to guard against bias that is sufficient to flip the sign of the treatment effect. In other words, if the treatment effect is positive, we are not worried about bias that inflates the treatment effect. Instead, we are worried about bias that would make us more likely to detect a negative treatment effect. Interestingly, this contrasts with typical intuitions about promoting conservatism or a passing a "harder test" of a theoretical prediction/implication.

Remark 1. *When **testing a directional prediction** is the goal of a study, the within-subjects design is preferred whenever:*

$$b_s \geq -\frac{\sqrt{se_e} - \sqrt{se_w}}{\sqrt{se_e}}\tau \text{ if } \tau > 0 \text{ or } b_s \leq -\frac{\sqrt{se_e} - \sqrt{se_w}}{\sqrt{se_e}}\tau \text{ if } \tau < 0.$$

If else, the experimental design is preferred.

Now consider the case in which a researcher's goal is to accurately estimate a causal effect. Here, I suggest that the researcher select the design that minimizes mean-squared error. Mean-squared error can be decomposed into variance and (squared) bias, which allows us to formalize the tradeoff that we are weighing. In the following expression, b_p^2 (p for point estimate) refers to

[11] Because researchers may not always be targeting the ATE, I use τ to represent the causal estimand of interest.

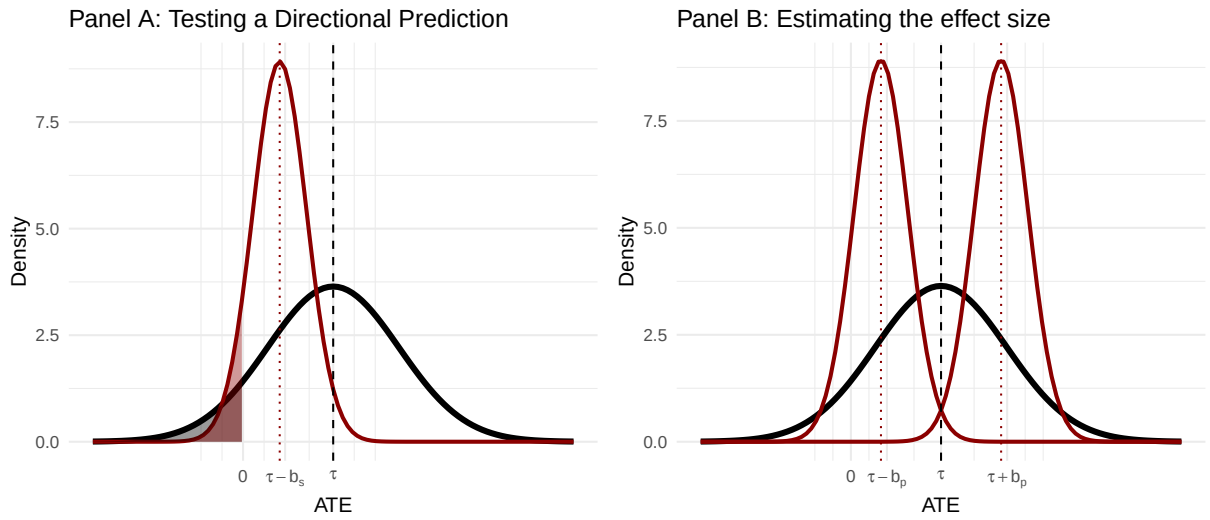

Figure 3: Both graphs plot sampling distributions at which a researcher would be indifferent between conducting an experiment and a within-subjects design. This indifference characterizes the maximum bias that a researcher would tolerate in a within-subjects design, b_s and b_p.

the bias of the within-subjects design at which we would be indifferent between the experiment and the within-subjects design.

$$\underbrace{se_e + 0}_{\text{Experiment}} = \underbrace{se_w + b_p^2}_{\text{Within-subjects design}}$$

Rearranging this expression yields Remark 2. Consistent with this remark, Panel B of Figure 3 plots the sampling distribution of the ATE from an experimental design in black along with the two within-subjects designs at which the researcher would be indifferent between the experimental and within-subjects designs. These distributions characterize the maximum magnitude of bias that a researcher would tolerate in exchange for variance reductions.

Remark 2. *When **estimating a treatment effect** is the goal of a study, the within-subjects design is preferred whenever:*

$$b_p \in \pm\sqrt{se_e^{ATE} - se_w^{ATE}}$$

Else, the experimental design is preferred.

It is useful to compare the properties of Remarks 1 and 2. First, note that the bias that we would

be willing to admit in the directional test depends on the treatment effect, τ. As τ becomes large, we would be willing to accept bias of a larger magnitude (in the opposite direction). In contrast, if we are simply trying to estimate a treatment effect, the degree of bias that we would be willing to accept does not depend on τ. Second, intuitively, both tests depend on the degree of efficiency gains of the within-subjects design. As this advantage of the within-subjects design becomes more pronounced relative to the between-subjects design, the magnitude of bias that we are willing to admit similarly increases.

Readers may ask whether the decision rules in Remarks 1 and 2 are unique to survey experiments. They are not. However, in other contexts, whether with observational data or with more time-intensive (field) experimental interventions, researchers typically do not have the choice between these designs. Where plausibility constrains the ways that researchers can ultimately induce or measure responses to treatment, these decision rules do not provide practicable guidance.

5 Practical Guidance: Choosing a Survey Design

This paper has explored the tradeoffs between experimental and between-subject designs and proposed decision rules between the two designs. How should applied researchers use this guidance when planning experiments? Here I propose two recommendations. The first emphasizes the prospective use of the decision rules and the second provides a tool for estimating the parameters of the decision rules through a hybrid design.

5.1 Prospective use of the decision rules

Experimentalists have correctly focused on improving research design in advance of conducting studies. Many of the metrics used to optimize research designs rely on making assumptions about the data generating process. For example, power calculation—whether conducted by analytical formulas or numerical simulation—relies on this type of assumption about the data. Blair et al. (2019) and Blair, Coppock, and Humphreys (2023) argue that use of simulation can lead researchers to evaluate many diagnosands beyond power to improve or understand the properties of their research designs.

The decision rules in Remarks 1 and 2 correspond straightforwardly to this exercise. The variance of estimators of the ATE for both can be calculated analytically (under some data generating processes) or estimated from simulated mock data (as in Blair et al. (2019) and Blair, Coppock, and Humphreys (2023)). Of course, these assumptions can be calibrated on the basis of existing or pilot data. Further, when researchers motivate their research design or hypotheses in a pre-analysis plan, they generally convey which of the two goals—testing a directional prediction or generating a point estimate—is more important.

What remains outside of common practice, then, is an assumption (or prediction) about the direction and magnitude of the bias induced by a within-subjects design. This could be justified by argument about which mechanisms for bias may be active given a set of interventions, outcomes, and the context. It could also be informed by measured estimates of, for example, priming effects or outcome stability in other contexts. While making assumptions about the magnitude of bias is not standard, these assumptions are arguably no more tenuous or fanciful than others invoked in any existing power calculation. Indeed, one can view the current reliance on experiments over within-subjects designs as an implicit assumption that bias is large. The decision rules that I articulate could be evaluated as part of the research design process for any survey-based study with an intervention.

5.2 Hybrid designs

It is also possible to design a study such that the decision rules can be used to guide analysis rather than prospective research design. To this end, we need a design that allows us to estimate three or four parameters, depending on the decision rule:

1. se_e: The variance of the experimental design
2. se_w: The variance of the within-subjects design
3. b: The bias of the within-subjects design
4. τ: The estimated ATE from the experimental design—only necessary for Decision Rule #1

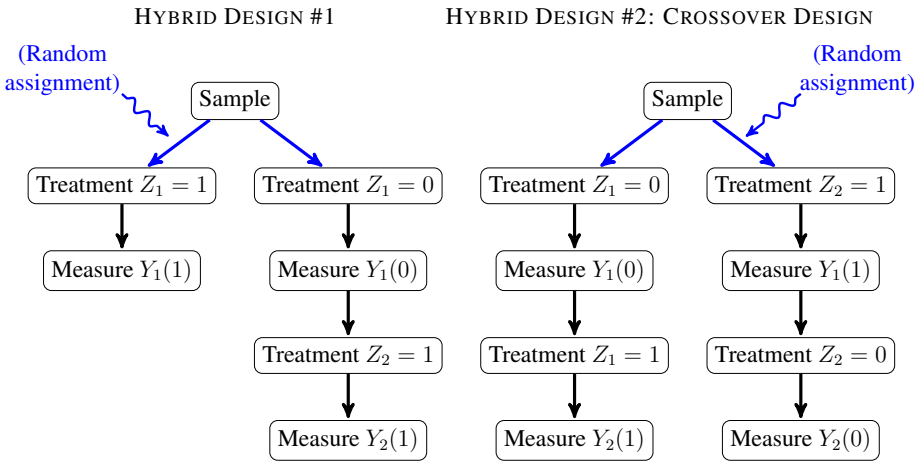

Figure 4: Simplified survey flow for proposed hybrid designs.

With measures of these quantities, it is straightforward to apply Remark 1 or 2. I provide a simplified survey flow for a hybrid design for any two-arm experiment in Figure 4. In hybrid design #1 it is infeasible to measure Treatment 0 after Treatment 1. This could be analogous to an intervention that reveals information.[12] Treatment 0 could be a pure control arm or placebo while Treatment 1 provides the informational stimulus. Since we cannot erase information from a subject's mind after revealing it, it makes little sense to pretend to measure control or placebo potential outcomes *after* the informational stimulus. To my knowledge, hybrid design #1 is original to this paper.[13]

In contrast, hybrid design #2, which is typically called a crossover design, we can provide subjects with both treatments in any order. We lose less information by treating all subjects with both interventions; we simply randomize the order. While the crossover design is not novel to this paper, the proposed method for using this design to adjudicate between the between- and within-subjects estimates of the treatment effect is novel to this paper and does not have a clear precedent among existing methods.

[12]This is also an appropriate design for vignette or question wording experiments that reveal some information (see Table A1) or emotion induction experiments that compare behavior in one mood/emotional state to a status quo (non-induced) emotion/mood.

[13]For comparisons of a treatment to a pure control condition, hybrid design #1 consists of a subset of two (of four) arms from the Solomon (1949) design. However, this property is not general to the case of two distinct treatments (e.g., a placebo and treatment condition), since the arm that measures responses to both treatments does not have an analogue in the Solomon design. As such, Hybrid design #1 should not be seen as a special case of the Solomon design.

Both designs nest a between-subjects experiment design. One estimator of the experimental ATE is:

$$\overline{Y}_{i1}(1|Z_1 = 1) - \overline{Y}_{i1}(0|Z_1 = 0)$$

This estimate is important because it helps us to measure the bias associated with the within-subjects design. Denoting the probability of assignment to treatment $Z_1 = 0$ as π, the within-subjects estimators of the ATE from each hybrid design are:

$$\overline{Y_{i2}(1|Z_1 = 0) - Y_{i1}(0|Z_1 = 0)} \quad \text{Hybrid design \#1}$$

$$\pi\overline{Y_{i2}(1|Z_1 = 0) - Y_{i1}(0|Z_1 = 0)} + (1-\pi)\overline{Y_{i1}(1|Z_1 = 1) - Y_{i2}(0|Z_1 = 1)} \quad \text{Hybrid design \#2}$$

By random assignment of Z_1, $E[Y_{i1}(1|Z_1 = 1)] = E[Y_{i1}(1|Z_1 = 0)]$ and $E[Y_{i1}(0|Z_1 = 1)] = E[Y_{i1}(0|Z_1 = 1)]$. By substitution, the estimators of the bias of each of the within-subjects designs are therefore:

$$\overline{Y}_{i2}(1|Z_1 = 0) - \overline{Y}_{i1}(1|Z_1 = 1) \quad \text{Hybrid design \#1}$$

$$\pi[\overline{Y}_{i2}(1|Z_1 = 0) - \overline{Y}_{i1}(1|Z_1 = 1)] - (1-\pi)[\overline{Y}_{i1}(0|Z_1 = 1) - \overline{Y}_{i2}(0|Z_1 = 1)] \quad \text{Hybrid design \#2}$$

It is useful to note that while the ATE in the within-subjects design is non-experimental, our estimator of the bias of the within-subjects design is experimental. The variance estimators of the ATE for both the experimental and within-subjects designs are straightforward. This design therefore yields estimates of all parameters of the decision rules in Remarks 1 and 2. Thus, hybrid design users can pre-commit to the following analytic strategy when they are uncomfortable committing to one design *ex-ante*:

1. Implement hybrid design #1 or #2.

2. Estimate $\widehat{se_e}$, $\widehat{se_w}$, $\widehat{b}$ (bias of the within-subjects design), and (as relevant) $\widehat{\tau}$.

3. Plug these estimates into the decision rule relevant to the substantive claim being made (Remark 1 or 2).

4. Select the experimental or within-subjects design based on the decision rule.

5. Implement a Bonferroni correction ($m = 2$) on the ATE estimate from the selected design to account for use of the same data for: (1) selecting the design and (2) estimating the ATE.

Several comments are in order. First users who seek to test a directional prediction using the decision rule in Remark 1 should use the experimental estimate of $\widehat{\tau}$. (The decision rule in Remark 2 does not depend on effect size.) Second, because this procedure uses the same data to select the design and estimate treatment effects, we must account for this dual use of the data when conducting inference. Following guidance on post-selection inference, I use a Bonferroni correction with $m = 2$ to account for the design selection and the estimation of treatment effects. This correction constructs $1 - \alpha$ confidence intervals by using the critical value given by $1 - \alpha/(2m) = 1 - \alpha/4$. Figure A4 shows that this correction is sufficient to achieve nominal coverage of $1 - \alpha$ in several simulations for both decision rules.[14] If one were to implement a hybrid design in a pilot survey in order to measure bias so that they could select a design for the ultimate survey, the final step (the correction for post-selection inference) would be unnecessary in analysis of data from the ultimate survey.

5.3 Properties of the hybrid design

The hybrid designs allow researchers to use their preferred decision rule without making strong assumptions about the underlying data generating process. To evaluate the performance of the proposed design, I conduct a number of simulations, elaborated in greater depth in Appendix A7. One important metric of performance evaluates the rate at which the implementation of the decision rules on simulated data agrees with the decision rules evaluated on the data generating process that produces the simulated data. The left panel of Figure 5, examines a number of data generating processes without bias at varying treatment effect sizes. In the case of a data generating

[14] In the absence of this correction, coverage rates are non-trivially lower than $1 - \alpha$.

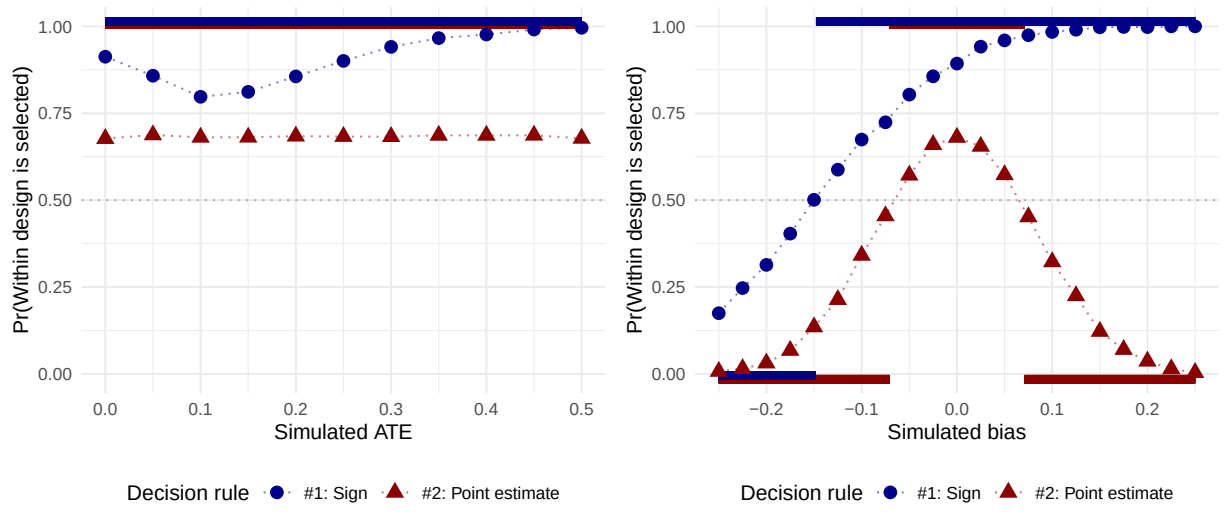

Figure 5: Probability of selection of the within-subjects design in two simulations of hybrid design #2 (the crossover design). The points and dashed lines report the results from the simulation. The thick, transparent lines represent the outcome of each decision rule implemented on the parameters of the data-generating process. The left simulation assumes $\text{Var}[Y_i(0)] = 1$, $\text{Var}[\tau_i] = 1$, $\text{Cov}[Y_i(0), \tau_i] = 0$, and $N = 1,000$. The right simulation assumes $E[\tau_i] = 0.25$, $\text{Var}[Y_i(0)] = 1$, $\text{Var}[\tau_i] = 1$, $\text{Cov}[Y_i(0), \tau_i] = 0$, and $N = 1,000$.

process without bias, both decision rules would select the within-subjects design. In Figure 5, we see that the simulation of the hybrid design selects the within-subjects design a majority of the time under both decision rules. With decision rule #1—for directional predictions—the within-subjects design is selected at a higher rate. This rate increases in effect size because decision rule #1 admits larger biases (in magnitude) as effect size increases. In contrast, decision rule #2—for point estimates—does not depend on effect size; this is evident from the simulation results that show that the probability that the within-subjects design is selected does not depend on the simulated effect size.

The simulations in the right panel of Figure 5 fix the effect size to 0.25 control group standard deviations and vary the bias of the within-subjects design from -0.25 to 0.25. As in the left panel, the thick, transparent lines show which design is selected under the specified data generating process. We see that, for both decision rules, the within-subjects design is selected with probability ≥ 0.5 whenever the within-subjects design is theoretically preferred given the data generating process. The probability that the within-subjects design is selected is < 0.5 whenever the experiment

is optimal.

Within these simulations, decision rule #1 is theoretically more likely to select the within-subjects design. It is important to clarify that this is not a general feature. To see this, compare the threshold b_s (as defined in Remark 1) to the threshold of the same sign, b_p^- or b_p^+ (defined in Remark 2). By straightforward examination of these expressions, when τ becomes sufficiently small and bias is present, there exist parameter sets for which decision rule #2 selects a within-subjects design and decision rule #1 selects an experiment.

It is important to consider the differences between hybrid design #1 and hybrid design #2. hybrid design #1 responds to logistical constraints by collecting less data than hybrid design #1. As a result, hybrid design #1 only collects within-subject data from half the sample. This reduces the likelihood of choosing the within-subjects design under either decision rule. This occurs because the efficiency gains of the within-subjects design are limited with half the number of observations as the experiment.

Finally, I have advocated the hybrid design as an agnostic method for choosing between the experimental and within-subjects designs. One might reasonably wonder if we could do better by *combining* the estimates using the hybrid design. Specifically, could we use the experimental estimate of bias to de-bias the more efficient within-subjects estimates? Unfortunately, Appendix A6 reveals that this approach cannot guarantee decreases in the variance relative to an experiment. This is because we rely on an experimental estimate of bias to de-bias estimates, which has the same problem as the experimental ATE estimate: lack of efficiency.

5.4 Comparing prospective use of decision rules to hybrid design

When should researchers use the decision rules prospectively relative to the hybrid design? Consider first the costs and benefits of the prospective use of decision rules. Suppose that a researcher uses the decision rules prospectively and chooses the *correct* design.[15] Relative to using the hybrid design, prospective use of the decision rules: (1) eliminates the efficiency loss from the Bonfer-

[15]By correct design, I mean that their selected design matches the design that decision rule would select when applied to the true (but unknown) data generating process.

roni correction and (2) avoids the possibility that the hybrid design selects the suboptimal analytic strategy. Further, when hybrid design #1 is the only feasible hybrid design, implementing a within-subjects analysis on the full sample generates more information than a within-subjects study on a random subset of the sample.

Now, consider the possibility that the researcher uses the decision rules prospectively and chooses the *wrong* design. If they incorrectly choose the within-subjects design, the efficiency gains do not compensate for the bias induced by the within-subjects design over the experiment. Moreover, unlike the hybrid design, this bias is unmeasured. If a researcher prospectively chooses the experimental design and is incorrect, they are likely to have less efficient estimates of treatment effects (relative to the hybrid design), thereby risking false negative inferences.

Clearly, then, there exists a tradeoff between prospective use of the decision rules and the hybrid design. Here, consideration of the substantive context of treatments and outcomes seems useful. Some treatments and outcomes are sufficiently tested that we have an idea of likely effect sizes and variance of common experimental estimators. These data can be used to reasonably calibrate τ and se_e. Due to the lack of within-subjects designs in current practice, though, we do not generally have estimates of the variance and associated standard errors of the within-subjects estimator. However, if outcomes of interest are measured in panel data in other studies/contexts, repeated measures may help to calibrate expectations about efficiency gain from the within-subjects design (and thus se_w). With reasonable calibrations of τ, se_e, and se_w, researchers can assess the range of biases that would support the adoption of either design. This discussion suggests that for well-studied treatment/outcome combinations, prospective use of the decision rules may be preferable. In contrast, when researchers have no idea what values these parameters might take, the hybrid design frees the researcher from making arbitrary assumptions. This means that in survey research on relatively less-studied interventions or outcomes, the hybrid design may be preferred.

5.5 Practical considerations: survey costs

To this point, I have largely ignored practical considerations in favor of a discussion of the potential for learning about treatment effects. But researchers are generally budget constrained. To this end,

a within-subjects design (1) requires a smaller sample sample size for equivalent power; but (2) will require that each respondent is exposed to more treatments and answers more questions (to collect outcomes under more treatment conditions). Discussions with five survey firms in anticipation of the original survey in Section 6 suggests a common pricing structure for survey research:

1. A fixed cost per respondent. The fixed cost depends on the size of the population as a proportion of a vendor's subject pool.

2. A variable cost per question or per minute of (anticipated) response time. This cost takes different functional forms, but was concave for all firms, meaning that the marginal cost of an additional minute or question weakly decreases as the survey becomes longer.

Ultimately, then, the budgetary implications of the tradeoff in these designs depends on whether the savings from sample size reductions afforded by the within-subjects design offset the increase in survey length. In the absence of lengthy treatments or a large number of outcomes for each treatment, savings from having fewer respondents are likely to be greater than additional costs from longer surveys. When using the decision rules prospectively, it may be useful to budget both the experimental and within subjects designs. In general, it seems that for short- to moderate-length treatments with a small-to-moderate number of outcomes, these practical considerations will favor the within-subjects design. The savings of the within-subjects design will become more pronounced as a subject pool becomes more specific due to the increase in fixed costs per respondent increase.

In contrast, use of the hybrid design will likely increase survey costs moderately relative to an experiment. Because the experimental design nests within the hybrid design, reducing the sample size of the hybrid design below the sample size a researcher would choose (prospectively) for an experiment is not advisable. The hybrid design requires increasing survey length (for at least a subset of respondents) which will increase the variable costs of a survey (holding fixed other content). If treatments are relatively short and there are a limited number of outcomes, these costs can be limited relative to the total cost of the survey. For example, in the original application that

Application	Hybrid design	Treatment conditions	Outcome
1	#1	**0**: No informational op-ed. **1**: Respondents read op-ed on climate change	Z-score index of items on climate change, climate policy (see items in Table A3)
2	#2	**0**: "Generally speaking, do you think we're spending too much, too little or about the right amount on *assistance to the poor*?" **1**: "Generally speaking, do you think we're spending too much, too little or about the right amount on *welfare*?"	**1**: too little; **2**: about the right amount; **3**: too much

Table 5: Application #1 comes from an original survey. Application #2 comes from Clifford, Sheagley, and Piston (2021).

follows, moving from an experiment to hybrid design #1 increased survey costs by 13% for half the sample, thereby increasing the total survey cost by 6.5%.

6 Application of Hybrid Designs

I provide one application of each hybrid design to illustrate use of the hybrid design and decision rules proposed in this article. The first application is an original study of the effects of an op-ed on mass beliefs following Coppock, Ekins, and Kirby (2018). The second application is a re-analysis of a crossover design-based question wording experiment published by Clifford, Sheagley, and Piston (2021). Table 5 summarizes the mapping of the applications to the hybrid designs. Both applications highlight the utility of the hybrid designs. Moreover, results in both applications point to the advantages of the within-subjects design over survey experimental designs.

6.1 Hybrid design #1 application: The effect of an op-ed on climate beliefs

In the first application, I measure the effects of an op-ed warning of the dangers of climate change on beliefs about climate change. The pre-registered survey was fielded online to a nationally-representative sample of 516 United States adults in December 2024. Using simple random assignment, respondents were assigned to one of the two arms of hybrid design #1 with probability $1/2$. A balance table (Table A4) and omnibus balance test (Table A5) offer no evidence against the integrity of the random assignment. In the first condition (the left arm of the left panel of Figure 4), respondents read the op-ed and then answered four questions about their beliefs about climate change. These responses measure *treated* potential outcomes. In the second condition (the

right arm of the left panel of Figure 4), respondents initially answered four questions about climate change beliefs then read the op-ed and answered the same four questions. The pre-treatment responses measure *untreated* potential outcomes while the post-treatment responses measure *treated* potential outcomes (that might be affected by the repeated questioning in addition to the op-ed). Table A3 reports the specific outcome measures employed. The pre-specified outcome of interest is a Z-score index of these four measures of beliefs about climate change.

I note that this setting should represent a relatively hard case for the within-subjects design. First, we are constrained to hybrid design #1 because after treating subjects with the op-ed, a researcher cannot simply expunge this information. This means that we only have a within-subjects design from (approximately) half the sample. Moreover, to the extent that we expect that subjects to learn in a Bayesian manner in response to the information presented in the op-ed, we would expect a negative covariance between baseline potential outcomes ($Y_i(0)$) and individual treatment effects (τ_i).[16] This negative covariance limits the efficiency gains of the within-subjects design.

Figure 6 reports ATE estimates from the experimental and within-subjects designs in the left panel. Both estimates suggest that, on average, the op-ed increases beliefs about the dangers of climate change and the priority afforded to climate adaptation/mitigation efforts. These standardized effects—0.172 in the experiment and 0.139 in the within-subjects design—are somewhat smaller than analogous studies of the effects of op-eds on mass beliefs in other policy domains (e.g., Coppock, Ekins, and Kirby, 2018). The within-subjects design clearly produces a smaller point estimate, but the difference in these estimates—a measure of the bias of the within-subjects design—is not distinguishable from zero (see Table A6). The more striking difference is the marked precision gain from the within-subjects design. This is evident from the shorter confidence interval on the within-subjects estimate.

The center and right panels reveal that both decision rules select the within-subjects ATE estimate for this application. In other words, for both goals—testing a directional prediction or estimating a treatment effect—the variance reductions in the within-subjects design outweigh the

[16]Figure A8 shows that we observe the negative covariance consistent with Bayesian updating in the data from the within-subjects arm.

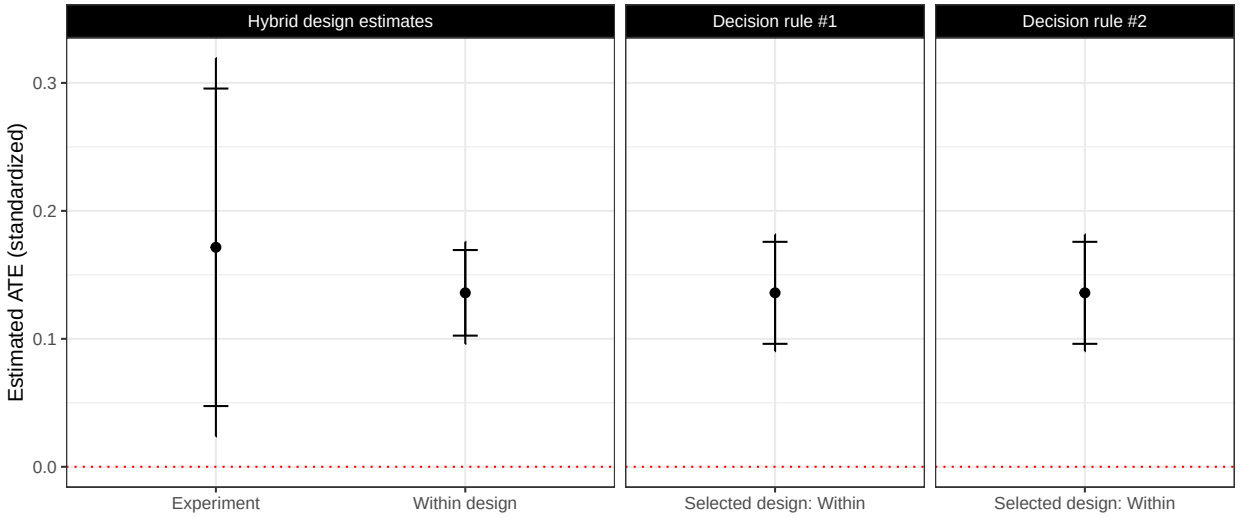

Figure 6: Estimates of the ATE using both designs in left panel for Application #1. The center and right panels show the application of each decision rule. Vertical (horizontal) segments denote 95% (90%) confidence intervals. See Table A6 for regression results used in the construction of this plot.

estimated bias induced by the repeated questions. The confidence intervals in the center and right panels are slightly longer than those on the within-subjects ATE estimate in the left panel. This is a consequence of the use of Bonferroni correction for the use of the same data for design selection and treatment effect estimation.

6.2 Hybrid design #2 application: Re-analysis of a question wording experiment

While the hybrid survey design in Application #1 is new (to the best of my knowledge), it is possible to use the decision rule with some existing within-subjects surveys. To this end, I reanalyze data from Clifford, Sheagley, and Piston (2021) following the methods developed in this paper. Substantively, this application is a question wording experiment, which represents an easy case for the within-subjects design. It compares support for "welfare" to support for "assistance to the poor." Figure A9 depicts the design of the applicable arms of the Clifford, Sheagley, and Piston (2021) study.

In this re-analysis, the within-subjects design does not appear to induce appreciable bias. The estimated ATE of "welfare" relative to "assistance to the poor" on public support for social spend-

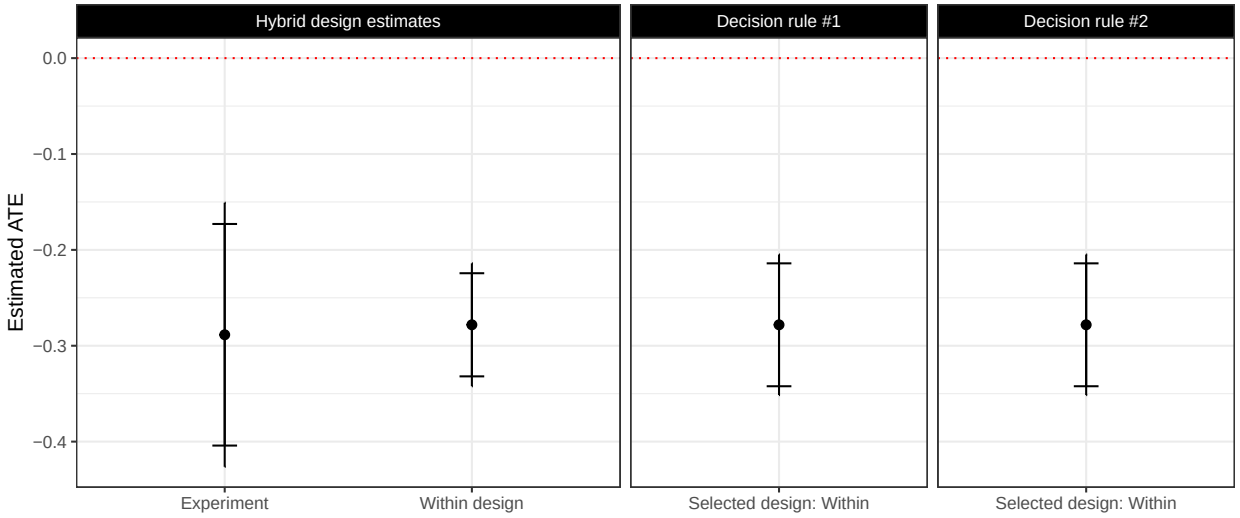

Figure 7: Estimates of the ATE using both designs in left panel for Application #2. The center and right panels show the application of each decision rule. Vertical (horizontal) segments denote 95% (90%) confidence intervals. See Table A7 for regression results used in the construction of this plot.

ing is -0.289 (95% CI: [-0.427, -0.151]) in the experiment and -0.278 (95% CI: [-0.342, -0.213]) in the within-subjects design. The similarity of these estimates suggests that the within subjects design does not induce appreciable bias in this application. Logically, these treatments are relatively inconspicuous and short which may well limit the possibility of priming, demand effects and respondent fatigue as sources as bias. As in the first application, Figure 7 reveals a stark efficiency gain from the within-subjects design, however. If attitudes toward welfare and support for the poor are correlated (e.g., due to preferences toward state intervention), we would expect substantial efficiency gains from the within-subjects comparison. Unsurprisingly, both decision rules select the within-subjects design in this case, as is evident in the center and right panels. Again, the slightly longer confidence intervals in these panels (relative to the raw ATE estimate from the within-subjects design) reflect the Bonferroni correction.

7 Conclusion

In this paper, I argue that by relying so heavily on survey experiments, social scientists may be unwittingly—and counterproductively—manufacturing missingness. This occurs when researchers

opt to measure only one potential outcome per subject in order to facilitate unbiased estimation of a causal effect. But in so doing, they sacrifice a substantial amount of information about individual responses to treatment. I argue that we can learn more about social and political phenomena by adopting a more pragmatic approach to the bias-variance tradeoff that is associated with the choice of experimentation versus non-experimental comparisons in within-subjects designs. The flexibility of surveys as a means for delivering treatments and measuring responses means that multiple designs are available. In this context, a commitment to design-based research compels us to consider the merits of designs beyond standard between-subjects experiments.

This paper makes two central contributions to survey research in the social sciences. First, I formulate the first non-heuristic decision rules that facilitate comparison of experiments to within-subjects design. In so doing, I make an important distinction between research intended to test directional effects versus generate point estimates of quantities of interest. This distinction is central to applied research but is rarely developed as a consideration in the methodological literature (but see Slough and Tyson, 2025). Second, for researchers who seek to let the data dictate an optimal design for intervention-oriented surveys, I propose two hybrid designs that facilitate use of the decision rules. Two applications attest to the wide applicability of these designs and suggest that a within-subjects design may be selected over a traditional survey experimental design in a range of circumstances due to their efficiency gains.

In sum, I provide new tools to improve research design for intervention-oriented surveys. These considerations, in conjunction with the near monopoly of between-subjects survey experiments, suggest that holding fixed the number of surveys, political scientists would be well served to conduct fewer between-subjects survey experiments in favor of within-subjects designs.

References

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto. 2019. "Beyond the breaking point? Survey satisficing in conjoint experiments." *Political Science Research and Methods* 9 (1): 53–71.

Berinsky, Adam J., Michele F. Margolis, and Michael W. Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58 (3): 739–753.

Blair, Graeme, Alexander Coppock, and Macartan Humphreys. 2023. *Research Design in the Social Sciences: Declaration, Diagnosis, and Redesign*. Princeton, NJ: Princeton University Press.

Blair, Graeme, Alexander Coppock, and Margaret Moor. 2020. "When to Worry about Sensitivity Bias: A Social Referent Theory and Evidence from 30 Years of List Experiments." *American Political Science Review* 114 (4): 1297–1315.

Blair, Grame, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113 (3): 838–859.

Briggs, Ryan, Jonathan Mellon, Vincent Arel-Bundock, and Tim Larson. 2025. "We used LLMs to Track Methodological and Substantive Publication Patterns in Political Science and They Seem to do a Pretty Good Job." Working paper.

Brutger, Ryan, Joshua D. Kertzer, Jonathan Renshon, Dustin Tingley, and Chagai M. Weiss. 2022. "Abstraction and Detail in Experimental Design." *American Journal of Political Science* 67 (4): 979–995.

Charness, Gary, Uri Gneezy, and Michael A. Kuhn. 2012. "Experimental methods: Between-subject and within-subject design." *Journal of Economic Behavior and Organization* 81 (1): 1–8.

Clifford, Scott, Geoffrey Sheagley, and Spencer Piston. 2021. "Increasing Precision wihtout Altering Treatment Effects: Repeated Measures Designs in Survey Experiments." *American Political Science Review* 115 (3): 1048–1065.

Coppock, Alexander, Emily Ekins, and David Kirby. 2018. "The Long-lasting Effects of Newspaper Op-Eds on Public Opinion." *Quarterly Journal of Political Science* 13 (1): 59–87.

Ding, Peng, and Fan Li. 2018. "Causal Inference: A Missing Data Perspective." *Statistical Science* 33 (2): 214–237.

Doyen, Stéphane, Olivier Klein, Cora-Lise Pichon, and Axel Cleeremans. 2012. "Behavioral Priming: It's All in the Mind, but Whose Mind?" *Plos One* 7 (1): e29081.

Esterling, Kevin, David Brady, and Eric Schwitzgebel. 2024. "The Necessity of Construct and External Validity for Deductive Causal Inference." Working paper, available at.

Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2006. "The Logic of the Survey Experiment." *Political Analysis* 15 (1): 1–20.

Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2004. *Problems and Methods in the Study of Politics*. Number 12 New York: Cambridge University Press chapter The illusion of learning from observational research, pp. 251–273.

Gillies, Jennifer C.P., and David J.A. Dozois. 2021. "How long do mood induction procedure (MIP) primes really last? Implications for cognitive vulnerability research." *Journal of Affective Disorders* 292: 328–336.

Graham, Mathew H., and Milan W. Svolik. 2020. "Democracy in America? Partisanship, Polarization, and the Robustness of Support for Democracy in the United States." *American Political Science Review* 114 (2): 392–409.

Grossman, Guy, William Dinneen, and Carolina Torreblanca. 2025. "The Evolving Landscape of Political Science: Two Decades of Scholarship in a Growing Discipline." Working paper available at `https://carolina-torreblanca.github.io/files/papers/landscape/landscape.pdf`.

Guess, Andrew, and Alexander Coppock. 2020. "Does Counter-Attitudinal Information Cause Backlash? Results from Three Large Survey Experiments." *British Journal of Political Science* 50 (4): 1497–1515.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22 (1): 1–30.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 396 (940-960).

Huber, Gregory A, and John S. Lapinski. 2006. "The "Race Card" Revisited: Assessing Racial Priming in Policy Contests." *American Journal of Political Science* 50 (2): 421–440.

Huber, Gregory A, and John S. Lapinski. 2008. "Testing the Implicit-Explicit Model of Racialized Political Communication." *Perspectives on Politics* 6 (1): 125–134.

Jones, Byron, and Michael G. Kenward. 2014. *Design and Analysis of Cross-Over Trials*. 3rd ed. New York: Chapman and Hall.

Kertzer, Joshua D. 2022. "Re-Assessing Elite-Public Gaps in Political Behavior." *American Journal of Political Science* 66 (3): 539–553.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams, Štěpán Bahník, Michael J. Bernstein, Konrad Bocian, Mark J. Brandt, Beach Brooks, Claudia Chloe Brumbaugh, Zeynep Cemalcilar, Jesse Chandler, Winnee Cheong, William E. Davis, Thierry Devos, Matthew Eisner, Natalia Frankowska, David Furrow, Elisa Maria Galliani, Fred Hasselman, Joshua A. Hicks, James F. Hovermale, S. Jane Hunt, Jeffrey R. Huntsinger, Hans IJzerman, Melissa-Sue John, Jennifer A. Joy-Gaba, Heather Barry Kappes, Lacy E. Krueger, Jaime Kurtz,

Carmel A. Levitan, Robyn K. Mallett, Wendy L. Morris, Anthony J. Nelson, Jason A. Nier, Grant Packard, Ronaldo Pilati, Abraham M. Rutchick, Kathleen Schmidt, Jeanine L. Skorinko, Robert Smith, Troy G. Steiner, Justin Storbeck, Lyn M. Van Swol, Donna Thompson, A. E. van 't Veer, Leigh Ann Vaughn, Marek Vranka, Aaron L. Wichman, Julie A. Woodzicka, and Brian A. Nosek. 2014. "Investigating Variation in Replicability." *Social Psychology* 45 (3): 142–152.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Survey." *Applied Cognitive Psychology* 5 (1): 213–236.

List, John A. 2025. "The Experimentalist Looks Within: Toward an Understanding of Within-Subject Experimental Designs The Experimentalist Looks Within: Toward an Understanding of Within-Subject Experimental Designs The Experimentalist Looks Within: Toward an Understanding of Within-Subject Experimental Designs." NBER Working Paper 33456.

Mendelberg, Tali. 2008*a*. "Racial Priming: Issues in Research Design and Interpretation." *Perspectives on Politics* 6 (1): 135–140.

Mendelberg, Tali. 2008*b*. "Racial Priming Revived." *Perspectives on Politics* 6 (1): 109–123.

Mummolo, Jonathan, and Erik Peterson. 2019. "Demand Effects in Survey Experiments: An Empirical Assessment." *American Political Science Review* 113 (2): 517–529.

Offer-Westort, Molly, Alexander Coppock, and Donald P. Green. 2021. "Adaptive Experimental Design: Prospects and Applications in Political Science." *American Journal of Political Science* 65 (4): 826–844.

Orne, Martin T. 1962. "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characterstics and their Implications." *American Psychologist* 17 (11): 776–783.

Schwarz, Susanne, and Alexander Coppock. 2022. "What Have We Learned about Gender from Candidate Choice Experiments? A Meta-Analysis of Sixty-Seven Factorial Survey Experiments." *The Journal of Politics* 84 (2).

Shanks, David R., Ben R. Newell, Eun Hee Lee, Divya Balakrishnan, Lisa Ekelund, Zarus Cenac, Fragkiski Kavvadia, and Christopher Moore. 2013. "Priming Intelligent Behavior: An Elusive Phenomenon." *Plos One* 8 (4): e56515.

Slough, Tara, and Scott A. Tyson. 2024. *Evidence Accumulation and External Validity*. New York: Cambridge University Press.

Slough, Tara, and Scott A. Tyson. 2025. "Sign-Congruence, External Validity, and Replication." *Political Analysis* Forthcoming.

Sniderman, Paul M. 2012. *Cambridge Hndbook of Experimental Political Science*. Number 8 New York: Cambridge University Press chapter The Logic and Design of the Survey Experiment: An Autobiography of a Methodological Innovation.

Sniderman, Paul M., and Thomas Piazza. 1993. *The Scar of Race*. New York: Cambridge University Press.

Solomon, R.L. 1949. "An Extension of Control Group Design." *Psychological Bulletin* 46 (2): 137–150.

Tyler, Matthew, Justin Grimmer, and Sean Westwood. 2024. "A Statistical Framework to Engage the Problem of Disengaged Survey Respondents: Measuring Public Support for Partisan Violence." Working paper, Rice University.

Weingarten, Evan, Qijia Chen, Maxwell McAdams, Jessica Yi, Justin Helper, and Dolores Albarricin. 2016. "From Primed Concepts to Action: A Meta-Analysis of the Behavioral Effects of Incidentally-Presented Words." *Psychological Bulletin* 142 (5): 472–497.