Manufacturing Missingness:
A Hidden Cost of Survey Experiments
Appendix

Tara Slough*

June 26, 2025

Contents

*Associate Professor, New York University, tara.slough@nyu.edu

A1 Survey of Experiments: Supplemental Information

Figure 1 relies on data from submissions to the EGAP registry from 2015-2019 (left panel) and an annual experiments conference (right panel). In both datasets, I rely on classification of experiments by authors. In the EGAP data, I construct an upper and a lower bound for the share of survey experiments. The upper bound is simply the share of projects that report to use a "survey methodology." All survey experiments should be classified as using a "survey methodology." I anticipate many of these studies are survey experiments due to the stronger norm of pre-registration for experiments than other designs (e.g., a non-experimental survey). The lower bound is given by studies classified as using both a "survey methodology" and an "experimental design." This is obviously a subset of studies included in the upper bound and approximates a reasonable lower bound on the share of survey experiments.

The conference submission form asks authors to classify the design directly as a survey, lab, field, or natural experiment. Some papers include more than one component. If a paper purports to use multiple methods, I include it in multiple categories.

A2 Proof of Proposition 1

For each unit, i, the following potential outcomes are well defined and real-valued:

$$\begin{aligned} &Y_i(0) \\ &Y_i(1) = Y_i(0) + \tau_i, \end{aligned}$$

where τ_i is the individual treatment effect. I assume that $\text{Var}[Y_i(0)] > 0$ and/or $\text{Var}[\tau_i] > 0$. The following proof generalizes the probabilities of assignment that are assumed to be $1/2$ in the main text. Specifically, I assume that $N \geq 4$ units of which m are assigned to treatment, where $2 \leq m \leq N-2$.[1] The standard errors associated with each ATE estimator in (3)-(4) are as follows.

$$\begin{aligned} \text{se}_w &= \sqrt{\frac{\text{Var}[\tau_i]}{N}} \\ \text{se}_e &= \sqrt{\frac{\text{Var}[Y_i(0)] + \text{Var}[\tau_i] + 2\text{Cov}[Y_i(0), \tau_i]}{m} + \frac{\text{Var}[Y_i(0)]}{(N-m)}} \\ &= \sqrt{\frac{(N-m)(\text{Var}[Y_i(0)] + \text{Var}[\tau_i] + 2\text{Cov}[Y_i(0), \tau_i])}{m(N-m)} + \frac{m\text{Var}[Y_i(0)]}{m(N-m)}} \\ &= \sqrt{\frac{N\text{Var}[Y_i(0)] + (N-m)(\text{Var}[\tau_i] + 2\text{Cov}[Y_i(0), \tau_i])}{m(N-m)}} \end{aligned}$$

[1] I assume that there are at least 2 units per treatment am so that the variance is estimable in each arm.

It is straightforward to show that $se_e > se_w$ when $\text{Var}[Y_i(0)] > 0$ and/or $\text{Var}[\tau_i] > 0$. Note that $\min \text{Cov}[Y_i(0), \tau_i] = -\sqrt{\text{Var}[Y_i(0)]}\sqrt{\text{Var}[\tau_i]}$. Plugging in this minimum covariance, $se_e > se_w$ iff:

$$\sqrt{\frac{N\text{Var}[Y_i(0)] + (N-m)(\text{Var}[\tau_i] - 2\sqrt{\text{Var}[Y_i(0)]}\sqrt{\text{Var}[\tau_i]})}{m(N-m)}} > \sqrt{\frac{\text{Var}[\tau_i]}{N}}$$

$$N\left(N\text{Var}[Y_i(0)] + (N-m)(\text{Var}[\tau_i] - 2\sqrt{\text{Var}[Y_i(0)]}\sqrt{\text{Var}[\tau_i]})\right) > m(N-m)\left(\text{Var}[\tau_i]\right)$$

$$N^2\text{Var}[Y_i(0)] + (N-m)^2\text{Var}[\tau_i] - 2N\sqrt{\text{Var}[Y_i(0)]}\sqrt{\text{Var}[\tau_i]} > 0$$

Recall that $m \leq N-2$ which implies that $(N-m)^2 \geq 4$. Since the left hand side of the inequality is increasing in $(N-m)^2$, I substitute $(N-m)^2 = 4$ and show that the the above inequality holds at this minimum.

$$N^2\text{Var}[Y_i(0)] + 4\text{Var}[\tau_i] - 2N\sqrt{\text{Var}[Y_i(0)]}\sqrt{\text{Var}[\tau_i]} \geq (N\sqrt{\text{Var}[Y_i(0)]} - 2\sqrt{\text{Var}[\tau_i]})^2 > 0$$

Therefore, this inequality must always hold when $\text{Cov}[Y_i(0), \tau_i] = \min \text{Cov}[Y_i(0), \tau_i]$. By inspection, se_e is increasing in $\text{Cov}[Y_i(0), \tau_i]$ whereas se_w is constant in $\text{Cov}[Y_i(0), \tau_i]$. Thus, if $\text{se}_e > \text{se}_w$ when $\text{Cov}[Y_i(0), \tau_i]$ is at its minimum, it must be the case that $\text{se}_e > \text{se}_w$ for any $\text{Cov}[Y_i(0), \tau_i]$. ■

A3 Classification of Treatments in Survey Experiments

Table 3 describes five types of treatments in survey experiments from an original hand-coding of the survey experiments assembled by Clifford, Sheagley, and Piston (2021). Here, I provide concrete examples of each type of experiment and document the overlap between multiple classes of experiments. The following list provides examples of each type of treatment.

1. Vignette treatment: Boas, Hidalgo, and Melo (2019) (Survey experiment)
    - Treatment arm 1: Vignette about hypothetical "Mayor Carlos" in a municipality like yours that includes sentence: "A State Accounts Court rejected the accounts of Mayor Carlos in 2013 because it found serious problems in the administration of the budget."
    - Treatment arm 2: Vignette about hypothetical "Mayor Carlos" in a municipality like yours with no information about accounts.
2. Information treatment: Christenson and Kriner (2019) (Experiment #1)
    - Treatment arm 1: Respondents read: "The current Congress has been one of the most obstructionist on record and is near historic lows in terms of its legislative productivity. Congress has failed to act on many of the most important issues facing the country. As a result of this congressional inaction, President Obama has aggressively used unilateral executive power to pursue his priorities in both foreign and domestic policy."
    - Treatment arm 2: Respondents read: "President Obama has aggressively used unilateral executive power to pursue his priorities in both foreign and domestic policy."

Type	Proportion	Proportion overlap with second category				
		Vignette	Information	Question wording	Incentives	Priming/induction
Vignette	53.8%	–	28.6%	2.9%	0%	0%
Information	40%	38.5%	–	11.5%	7.7%	0%
Question wording	15.4%	10%	30%	–	0%	0%
Incentives	7.8%	0%	40%	0%	–	0%
Priming/induction	6.1%	0%	0%	0%	0%	–

Table A1: Overlap between treatment classifications. The proportion of studies is reproduced from Table 3. The overlap proportions are conditional on the experimental type. For example, among vignette treatments, 28.6% of experiments are also classified as information treatments and 2.9% are also classified as question wording treatments.

3. Question wording treatment: Pérez and Tavits (2019). Subjects were asked their agreement with:
    - Treatment arm 1: "Calling on party leaders to encourage more women to run for office, a proposal that about 80% of the people in Estonia favor."
    - Treatment arm 2: "Calling on party leaders to encourage more women to run for office."
4. Incentives treatment: DeScioli, Shaw, and Delton (2018) (Experiment #1)
    - Treatment arm 1: *Risk pooling*: Respondent decides whether to keep endowment or invest in a risky lottery. Risk is pooled over four players' investment decisions and payoffs are shared across the four players.
    - Treatment arm 2: *Individual risk*: Respondent decides whether to keep own endowment or invest in a risky lottery. Payoff is based on their own choice (and, if they invest, the lottery outcome).
5. Priming/induction of emotion: Banks and Hicks (2016)
    - Treatment arm 1: Subjects see a picture of someone who is angry, are asked to write about something that made them angry.
    - Treatment arm 2: Subjects see a picture of someone who is afraid, are asked to write about what makes them afraid.
    - Treatment arm 3: Subjects asked to write about something that makes them relaxed.

Table A1 reports the overlap in the classifications of treatments. It suggests that informational treatments are commonly combined with vignettes, question wording, and incentive treatments. There is less overlap outside of informational treatments.

A4 Sample Size for the Difference in CATEs

The main text focuses on the average treatment effect. Many survey experiments also seek to measure heterogeneity in treatment effects, often by comparing conditional average treatment effects for multiple subgroups. It is well known that for a fixed effect and sample size, power to detect a difference in CATEs. This section provides analysis of the power of the experimental and within designs to detect differences in CATEs for two subgroups. Since the purpose of this analysis is illustration, I focus on the case in which a

binary treatment is assigned with probability $1/2$. I will consider a binary covariate, $X_i \in \{0, 1\}$ that takes the value 1 with probability $1/2$ which is independent of $Y_i(0)$, $X_i \perp Y_i(0)$. This case simplifies the standard error formulas substantially while illustrating properties of the two designs for measuring differences in CATES.

We will consider the following data-generating process:

$$Y_i(1) = Y_i(0) + \tau_i X_i + \varepsilon_i,$$

where $Y_i(1) \in \mathbb{R}$ and $\tau_i \in \mathbb{R}$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. It is straightforward to see that the difference in CATEs is:

$$\begin{aligned} CATE(X_i = 1) - CATE(X_i = 0) &= E[Y_i(1) - Y_i(0) \mid X_i = 1] - E[Y_i(1) - Y_i(0) \mid X_i = 0] \\ &= E[\tau_i + \varepsilon_i] - E[\varepsilon_i] \\ &= E[\tau_i]. \end{aligned}$$

Considering the following estimators of the difference in CATEs under each design:

$$\begin{aligned} &\text{Within design:} && \overline{[Y_i(1) - Y_i(0) \mid X_i = 1]} - \overline{[Y_i(1) - Y_i(0) \mid X_i = 0]} \\ &\text{Experiment:} && \overline{Y_i(1|X_i = 1)} - \overline{Y_i(0|X_i = 1)} - (\overline{Y_i(1|X_i = 0)} - \overline{Y_i(0|X_i = 0)}) \end{aligned}$$

In the within design, the standard error is given by:

$$\begin{aligned} se_w &= \sqrt{\frac{\text{Var}[\tau_i] + \sigma^2}{N/2} + \frac{\sigma^2}{N/2}} \\ &= \sqrt{\frac{2\text{Var}[\tau_i] + 4\sigma^2}{N}}. \end{aligned}$$

In the experimental design, the standard error is given by:

$$\begin{aligned} se_e &= \sqrt{\frac{\text{Var}[Y_i(0)] + \text{Var}[\tau_i] + \text{Cov}[Y_i(0), \tau_i] + \sigma^2}{N/4} + \frac{\text{Var}[Y_i(0)]}{N/4} + \frac{\text{Var}[Y_i(0)] + \sigma^2}{N/4} + \frac{\text{Var}[Y_i(0)]}{N/4}} \\ &= \sqrt{\frac{4(\text{Var}[\tau_i] + \text{Cov}[Y_i(0), \tau_i]) + 8\sigma^2 + 16\text{Var}[Y_i(0)]}{N}} \end{aligned}$$

Solving for the minimum sample size necessary to achieve 80% power, as in the manuscript, we have:

$$\begin{aligned} &\text{Within design:} && \Phi^{-1}(1 - \beta, 1.96, 1) = \frac{E[\tau_i]}{\sqrt{\frac{2\text{Var}[\tau_i] + 4\sigma^2}{N_w}}} \\ &&& \Rightarrow N_w = \frac{2\text{Var}[\tau_i] + 4\sigma^2}{(\frac{E[\tau_i]}{\Phi^{-1}(1-\beta, 1.96, 1)})^2} \\ &\text{Experiment:} && \Phi^{-1}(1 - \beta, 1.96, 1) = \frac{E[\tau_i]}{\sqrt{\frac{4(\text{Var}[\tau_i] + \text{Cov}[Y_i(0), \tau_i]) + 8\sigma^2 + 16\text{Var}[Y_i(0)]}{N_e}}} \\ &&& \Rightarrow N_e = \frac{4(\text{Var}[\tau_i] + \text{Cov}[Y_i(0), \tau_i]) + 8\sigma^2 + 16\text{Var}[Y_i(0)]}{(\frac{E[\tau_i]}{\Phi^{-1}(1-\beta, 1.96, 1)})^2} \end{aligned}$$

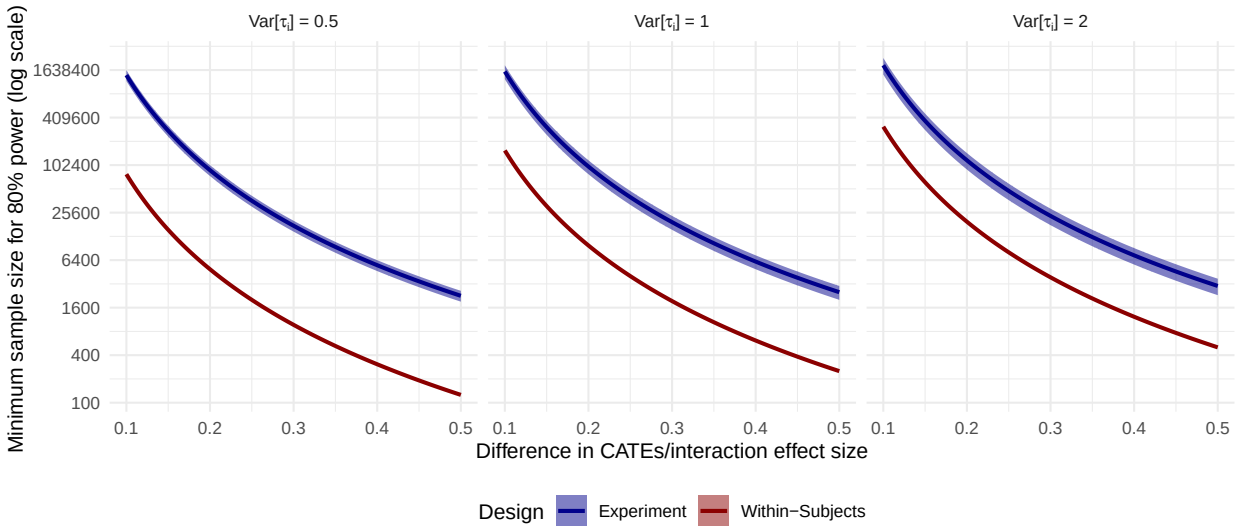

Figure A1: Minimum sample sizes for 80% power under different parametric assumptions. All simulations assume that $\text{Var}[Y_i(0)] = 1$ and $\sigma = 0$. The three panels depict different values of $\text{Var}[\tau_i]$. The intervals for the experimental design plot the minimum samples under the minimum and maximum covariances between τ_i and $Y_i(0)$, i.e., $\text{Cov}[Y_i(0), \tau_i] = \pm\sqrt{\text{Var}[Y_i(0)]}\sqrt{\text{Var}[\tau_i]}$, whereas the line plots the minimum sample size when $Y_i(0) \perp \tau_i$.

Figure A1 plots these minimum sample sizes under various parametric assumptions. These plots fix $\text{Var}[Y_i(0)] = 1$ and $\sigma = 0$ for comparability with the main text. The simulated difference in CATES, $E[\tau_i]$, varies along the x-axis. The ribbons/intervals for the experimental design depict the maximum range of $\text{Cov}[Y_i(0), \tau_i]$. The panels vary $\text{Var}[\tau_i]$. All plots reflect the substantial efficiency gains of the within-subjects design.

A5 No Evidence of Ordering Effects: Examination of Conjoint Experiments

I examine a set of twelve recent candidate choice conjoint surveys that were compiled by Schwarz and Coppock (2022).[2] In each of these conjoint surveys, respondents evaluate two profiles—potential candidates for office—with randomized characteristics. Respondents must choose the candidate that they prefer. These conjoint surveys contain different attributes (e.g., party, gender, or education) and levels within those attributes. Crucial for the present assessment of ordering effects in the within-subjects design, all of the included surveys ask respondents to evaluate more than one set of profiles. The number of sets of profiles ranges from 3 in Arnesen, Duell, and Johannesson (2019) to 30 Bansak et al. (2019). The order of of the profile serves as the (randomized) treatment of interest in this analysis.

To evaluate whether ordering affects responses (in the aggregate), I evaluate how ratings of conjoint profiles vary is affected by the (randomized) order in which they are presented to subjects. Consider a conjoint task in which respondents select one of two candidate profiles. We examine how a given level ("Republican") of an attribute ("party") varies as a function of the order of the profiles (e.g., first set, second set, etc.) on which that level appears. To do so, I estimate the following equations using OLS:

[2] I thank the authors for sharing systematized raw data from these experiments.

Constitutent study citation	Study	Sets of profiles	Attribute levels
Arnesen, Duell, and Johannesson (2019)		3	24
Atkeson and Hamel (2020)		3	16
Bansak et al. (2019)	1	30	57
Bansak et al. (2019)	2	30	57
Blackman and Jackson (2021)	1	4	36
Blackman and Jackson (2021)	2	5	31
Carey et al. (2022)	1	10	15
Carey et al. (2022)	2	10	15
Hainmueller, Hopkins, and Yamamoto (2014)		5	40
Kirkland and Coppock (2018)	1	5	24
Kirkland and Coppock (2018)	2	5	27
Senninger and Bischof (2023)		5	24
Total			366

Table A2: Study refers to the number of the study in the original article (for articles which report more than one distinct conjoint study).

$$\text{Profile selected}_{ital} = \beta \text{Profile set}_{ital} + \boldsymbol{\gamma}_i \qquad \text{for Attribute} = a,\ \text{Level} = l \qquad (1)$$

$$\text{Profile selected}_{ital} = \delta_0 + \sum_{t=2}^{T} \delta_t I[\text{Profile set}_{ital} = t] + \boldsymbol{\gamma}_i \qquad \text{for Attribute} = a,\ \text{Level} = l. \qquad (2)$$

The dependent variable, Profile selected$_{ital}$ is an indicator that takes the value 1 when a respondent selected a given profile (from a set of two). The treatment is the order in which a given profile was viewed, e.g. "1" for the first set, "2" for the second set, etc. The first specification, (1), is modeled linearly and Profile set$_{ital}$ ranges from 1 to the number of profiles (T). Ths second specification models the order of treatment more flexibility by estimating treatment effects (relative to the first profile, $t = 1$) for each t. In the main specifications, I include individual fixed effects $\boldsymbol{\gamma}_i$.[3] Note that I estimate (1)-(2) separately for each attribute and level. This means that we have $\sum_s a_s l_{sa}$ tests of each null hypothesis, in which a_s is the number of attributes in study s and l_{sa} is the number of levels in attribute a in study s.

Within these specifications, I test two null hypotheses that examine whether the probability that profile with a given attribute level is selected varies in the order of the profile, t. First, I test a null hypothesis that $\beta = 0$ in (1). Rejection of this null hypothesis suggests that the probability of selection increases (or decreases) in the order of the profile. Second, I test a null hypothesis that $\delta_T = 0$. Rejection of this null hypothesis suggests that the probability of selection is different in the first and last profiles viewed. Figure A2 illustrates these tests for one example level, "Republican," of the attribute "party" in Carey et al. (2022). This specific example suggests noisy increases in the probability that a Republican profile is selected over the 10 profiles given to respondents. The "marginal means" are depicted for each set of profiles (t). The first test evaluates the slope of the dot-dashed line (given by β). In this test, $p = 0.09$. The second test evaluates whether the marginal means in the first and last periods are equivalent, thus comparing the dark red estimates (with

[3] I include specifications without respondent fixed effects for robustness.

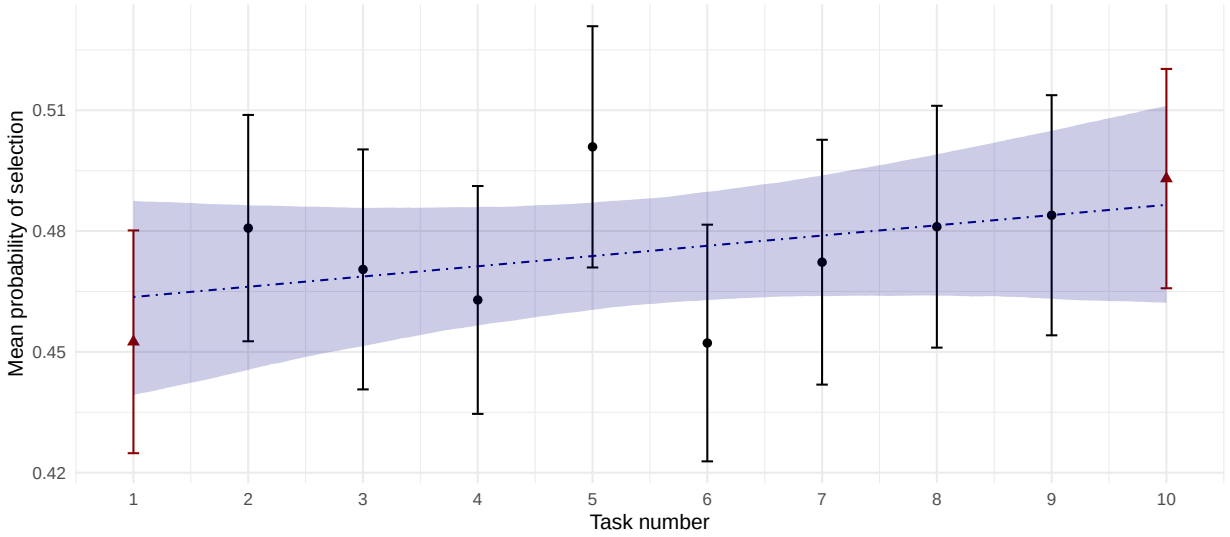

Figure A2: Illustration of comparisons of interest in the re-analysis of conjoint experiments. This graph reports marginal means (probabilities of profile selection) for "Republican" candidates from Carey et al. (2022). For this specific attribute level, the p-value of slope of the dot-dashed line is 0.09 and the p-value of the first-to-last task comparison is 0.04.

triangles). In this test, $p = 0.04$.

Whereas these tests are marginally significant, Figure A3 reports that this finding is rare. I plot the p-values from 366 tests of both null hypotheses. The left panel reports the p-values for the null that $\beta = 0$ (in (1)) and the right panel reports the p-values for the null that $\delta_T = 0$ (in (2)). The black line plots the empirical CDF of p-values from the specifications with respondent fixed-effects and the grey line plots the empirical CDF of p-values from tests without respondent fixed effects. If respondents were responding to profiles systematically differently over time, we would expect the ECDF curves to exceed the 45-degree line (in red). They do not. If anything, we observe *higher* p-values than we would expect from chance alone for both specifications and both tests. This is likely due to the lack of independence between responses to different attributes induced by the forced-choice decision rule.[4]

A6 Combining Experimental and Within-Subjects Estimates?

The principal tradeoff that I describe suggests that experiments provide unbiased estimators of the ATE (or other aggregate causal estimands) whereas the within-subjects design may admit bias but provides estiamtors that generally offer efficiency gains. To this end, one might ask whether, using the hybrid design, we can use the experimental estimate of the ATE to de-bias the within-subjects estimate fo the ATE. While this is, of course, possible. I show that this de-biasing strategy does not guarantee efficiency gains relative to the experiment. This occurs because the estimate of the bias used to in the de-biasing comes from the

[4]This lack of independence should create a *favorable* environment for detecting ordering effects—if we were to observe strong ordering effects for any level/attribute, we should also pick up compensating ordering effects in the opposite direction for some other level(s)/attribute(s).

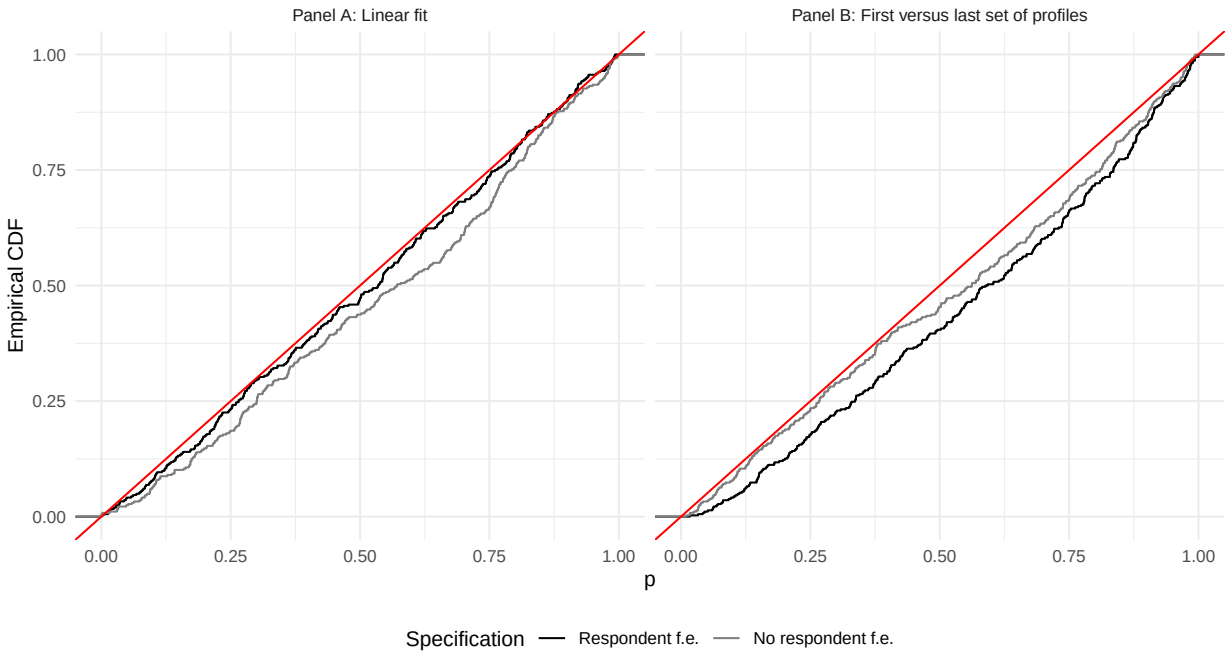

Figure A3: Each ECDF includes p-values from tests of ordering effects of 366 attribute levels from 12 conjoint studies that vary candidate attributes.

experimental design, with all standard efficiency limitations.

Suppose that we implement Hybrid Design #1. Subjects are randomized to each arm $O_i \in \{0, 1\}$ with probability $1/2$. If $O_i = 0$, we observe:

$$Y_{i1}(1) = Y_{i1}(0) + \underbrace{\tau_i}_{ITE}$$

If $O_i = 1$, we observe:

$$\begin{aligned} Y_{i1}(0) &= Y_{i1}(0) \\ Y_{i2}(1) &= Y_{i1}(0) + \underbrace{\tau_i}_{ITE} + \underbrace{\delta_i}_{\text{Artifact}} \end{aligned}$$

Note that the artifact is the source of (potential) bias in the within design. First, consider the estimators of the ATE under each design:

$$\begin{aligned} ATE_e &= \overline{Y_{i1}(1|O_i = 0)} - \overline{Y_{i1}(0|O_i = 1)} \\ ATE_w &= \overline{[Y_{i2}(1) - Y_{i1}(0)|O_i = 1]} \end{aligned}$$

Now note the following experimental estimator of the bias of the within-subjects design:

$$b_w = \overline{Y_{i2}(1|O_i = 1)} - \overline{Y_{i1}(1|O_i = 0)}$$

Now consider the de-biased estimator of ATE of the within-subjects design:

$$\begin{aligned} ATE_w^{db} &= ATE_w - b_w \\ &= \overline{[Y_{i2}(1) - Y_{i1}(0)|O_i = 1]} - \left(\overline{Y_{i2}(1|O_i = 1)} - \overline{Y_{i1}(1|O_i = 0)}\right) \end{aligned}$$

Second, the standard errors of the (baseline) ATE estimators are:

$$\begin{aligned} se_e &= \sqrt{\frac{\text{Var}[Y_i(0)] + \text{Var}[\tau_i] + 2\text{Cov}[Y_i(0), \tau_i]}{N/2} + \frac{\text{Var}[Y_i(0)]}{N/2}} \\ &= \sqrt{\frac{2\text{Var}[\tau_i] + 4\left(\text{Var}[Y_i(0)] + \text{Cov}[Y_i(0), \tau_i]\right)}{N}} \\ se_w &= \sqrt{\frac{\text{Var}[\tau_i] + \text{Var}[\delta_i] + 2\text{Cov}(\tau_i, \delta_i)}{N/2}} \\ &= \sqrt{\frac{2\text{Var}[\tau_i] + 2\text{Var}[\delta_i] + 4\text{Cov}(\tau_i, \delta_i)}{N}} \end{aligned}$$

The standard error of the bias estimator is:

$$\begin{aligned} se_{b_w} &= \sqrt{\frac{\text{Var}[Y_i(0)] + \text{Var}[\tau_i] + \text{Var}[\delta_i] + 2\text{Cov}[Y_i(0), \tau_i] + 2\text{Cov}[\tau_i, \delta_i] + 2\text{Cov}[Y_i(0), \delta_i]}{N/2} + \frac{\text{Var}[Y_i(0)] + \text{Var}[\tau_i] + 2\text{Cov}[Y_i(0), \tau_i]}{N/2}} \\ &= \sqrt{\frac{4\left(\text{Var}[Y_i(0)] + \text{Var}[\tau_i] + \text{Cov}[Y_i(0), \delta_i] + \text{Cov}[\tau_i, \delta_i]\right) + 2\text{Var}[\delta_i] + 8\text{Cov}[Y_i(0), \tau_i]}{N}} \end{aligned}$$

Finally the standard error of the de-biased ATE estimator is:

$$\begin{aligned} se_w^{\text{db}} &= \sqrt{\frac{2(\text{Var}[\tau_i] + \text{Var}[\delta_i]) + 4\text{Cov}[\tau_i, \delta_i] + 4\left(\text{Var}[Y_i(0)] + \text{Var}[\tau_i] + \text{Cov}[Y_i(0), \delta_i] + \text{Cov}[\tau_i, \delta_i]\right) + 2\text{Var}[\delta_i] + 8\text{Cov}[Y_i(0), \tau_i]}{N}} \\ &= \sqrt{\frac{6\text{Var}[\tau_i] + 4(\text{Var}[Y_i(0)] + \text{Var}[\delta_i] + \text{Cov}[Y_i(0), \delta_i]) + 8(\text{Cov}[Y_i(0), \tau_i] + \text{Cov}[\tau_i, \delta_i])}{N}} \end{aligned}$$

There are efficiency gains from the within estimator with debiasing relative to the experimental design if:

$$\begin{aligned} 2\left(3\text{Var}[\tau_i] + 2\text{Var}[Y_i(0)] + 2\text{Var}[\delta_i] + 2\text{Cov}[Y_i(0), \delta_i] + 4\text{Cov}[Y_i(0), \tau_i] + 4\text{Cov}[\tau_i, \delta_i]\right) &< 2\left(\text{Var}[\tau_i] + 2\text{Var}[Y_i(0)] + 2\text{Cov}[Y_i(0), \tau_i]\right) \\ 4\left(\text{Var}[\tau_i] + \text{Var}[\delta_i] + \text{Cov}[Y_i(0), \delta_i] + \text{Cov}[Y_i(0), \tau_i] + 2\text{Cov}[\tau_i, \delta_i]\right) &< 0 \end{aligned}$$

It is straightforward to see that this inequality does not hold if $\text{Var}[\tau_i] > 0$ and all covariance terms are zero (i.e., untreated potential outcomes, individual treatment effects, and artifacts are all independent). However, for an instance in which this does hold, suppose that treatment effects are homogeneous (e.g., $\text{Var}[\tau_i] = 0$). This implies that $\text{Cov}[Y_i(0), \tau_i] = 0$ and $\text{Cov}[\tau_i, \delta_i] = 0$. Setting $\text{Cov}[Y_i, \delta_i]$ to its theoretical minimum, $-\sqrt{\text{Var}[Y_i(0)]}\sqrt{\text{Var}[\delta_i]}$, then the inequality simplifies to:

$$\text{Var}[\delta_i] - \sqrt{\text{Var}[Y_i(0)]}\sqrt{\text{Var}[\delta_i]} < 0$$

This inequality is satisfied whenever $\text{Var}[Y_i(0)] > \text{Var}[\delta_i]$. These two cases show that a combined estimator that uses the experimental estimate of bias to debias the within-subjects ATE cannot guarantee efficiency gains over the experimental estimator of the ATE.

A7 Simulations investigating the properties of hybrid design

A7.1 Data generating process

To investigate the properties of the hybrid design as a means to select between the experimental and within-subjects designs, I employ a number of simulations. To conduct simulations, I generate data via the following DGP:

$$\begin{pmatrix} Y_{i1}(0) \\ \tau_i \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \begin{pmatrix} 1 & \text{Cov}[Y_{i1}(0), \tau_i] \\ \text{Cov}[Y_{i1}(0), \tau_i] & \text{Var}[\tau_i] \end{pmatrix} \right]$$

This implies that $\text{Var}[Y_{i1}(0)] = 1$, which is akin to standardizing the control potential outcome. I vary the following moments/parameters: $E[\tau_i]$, $\text{Cov}[Y_i(0), \tau_i]$, and $\text{Var}[\tau_i]$. I then generate a treatment indicator $Z_{i1} \in \{0, 1\}$ using complete random assignment, in which

$$\Pr(Z_{i1} = 1) = 1/2.$$

I allow for a bias of $b \in \mathbb{R}$ for the second treatment. This permits revelation of the following potential outcomes:

$$\begin{aligned} Y_{i1}(Z_{i1}) &= Y_{i1}(0) + Z_{i1}\tau_i \\ Y_{i2}(Z_{i2}) &= Y_{i1}(0) + Z_{i2}(\tau_i + b) \\ &= Y_{i1}(0) + (1 - Z_{i1})(\tau_i + b) \end{aligned}$$

This full DGP corresponds to Hybrid Design #2 since all subjects are treated with both treatments in a randomly-assigned order. To simulate Hybrid Design #1, note that in this design, $Y_{i2}(1|Z_{i1} = 0)$ is measured but $Y_{i2}(0|Z_{i1} = 1)$ is not.

A7.2 Simulations

The Monte Carlo simulations vary the following parameters:

- N: number of respondents
- $E[\tau_i]$: simulated ATE
- $\text{Var}[\tau_i]$: variance of treatment effects
- $\text{Cov}[Y_i(0), \tau_i]$: covariance of untreated potential outcomes (at $t = 1$) and treatment effects
- b: bias induced by within-subjects design

Note that the benchmark "correct" designs (e.g., in Figure 5) are calculated analytically using these parameters. Each of the reported simulation results is based on 3,000 iterations of the Monte Carlo simulations.

A7.3 Additional Results

Figure A4 reports the coverage of the (Bonferroni-corrected) 95% confidence intervals under Hybrid Designs #1 and #2. While these appear slightly conservative, naive confidence intervals (without the Bonferroni correction) consistently achieve coverage rates lower than 95%.

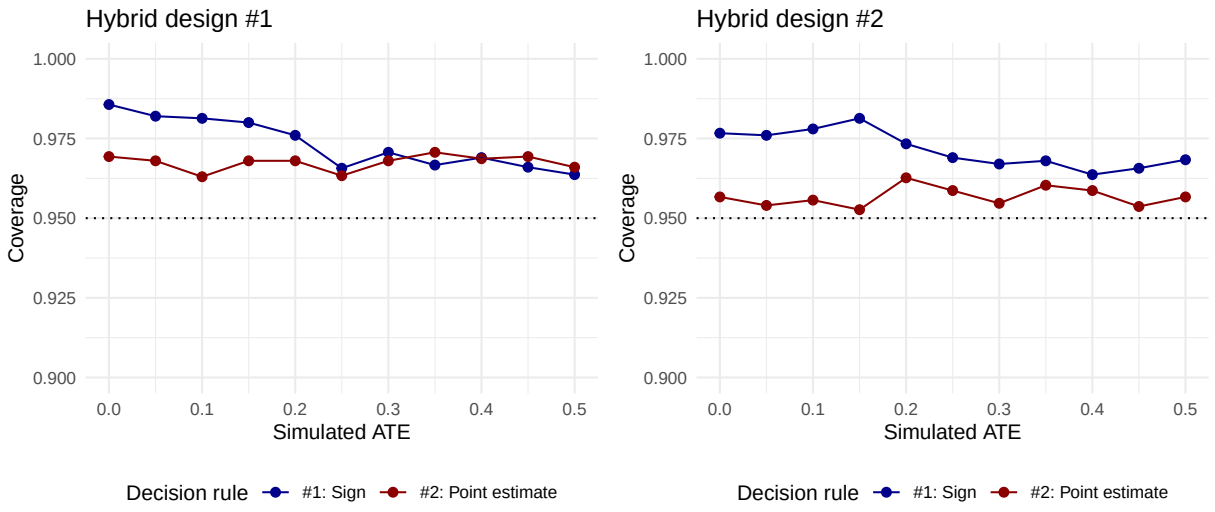

Figure A4: Coverage calculated from simulations varying effect size without the introduction of bias.

A8 Hybrid Design Applications

A8.1 Application #1: Responses to Climate Change Informational Op-Ed

A8.1.1 Design

I conduct an original survey with Hybrid Design #1 embedded. The design of the survey is depicted in Figure A5. I use simple random assignment to assign subjects to the two arms of the design. The informational treatment comes from the following op-ed: `https://www.nytimes.com/2023/10/13/opinion/climate-change-excessive-heat-2023.html`.

The outcome of interest is a Z-score index of the questions in Table A3. To construct the index, I reverse the order of questions such that increasing values of the constitutent outcome measures reflect agreement with the op-ed.

Tables A4-A5 provide two assessments of covariate balance across the two arms of Hybrid Design #2. Table A4 reports covariate means in each arm of the design and the p-values associated with differences in these means. There is not evidence of covariate imbalance. Table A5 provides an omnibus test of balance across the collected covariates. The p-values associated with the F-statistics (listed in the bottom of the table) suggest that covariates are not prognostic of treatment assignment. This is what we would expect given the random assignment of treatment.

A8.1.2 Ethical Practices Considerations

The original survey conducted for this application constitutes human subjects research. Below I describe the ethical practices and considerations relevant to this research, as articulated in APSA's *Principles and Guidance for Human Subjects Research*.

	Source	Outcome question	Response levels
1	CCES†	"Which of the following statements comes closest to your opinion about global climate change?"	(1) "Climate change is an extremely serious problem" . . . (5) Climate change is not a problem
2	Pew	"How much do you think human activity, such as the burning of fossil fuels, contributes to climate change?"	(1) A great deal . . . (4) Not at all
3	Pew	"How much of a priority should dealing with global climate change be for the president and Congress this year?"	(1) Top priority . . . (4) Shouldn't be done
4	Pew†	"Right now, which one of the following do you think should be the more important priority for addressing America's energy supply?"	(1) Developing alternative sources, such as wind, solar, and hydrogen technology; (2) Expanding exploration and production of oil, coal, and natural gas.

Table A3: Constituent measures in outcome index. Note revised measures. The † symbol indicates that question wording has been changed from the referent question in Pew or CCES.

	Assignment to belief measurement		Difference	
	Post-treatment	Pre- and post-treat.		
Indicator	Mean (Std. Dev.)	Mean (Std. Dev.)	T-statistic	p-value
Female	0.481 (0.501)	0.52 (0.501)	0.88	0.38
Under age 30	0.141 (0.349)	0.15 (0.357)	0.27	0.79
Age 30-40	0.176 (0.381)	0.209 (0.407)	0.95	0.34
Age 40-50	0.172 (0.378)	0.154 (0.361)	-0.56	0.58
Age 50-60	0.198 (0.4)	0.193 (0.395)	-0.16	0.87
Age 60+	0.313 (0.465)	0.295 (0.457)	-0.44	0.66
Democrat†	0.366 (0.483)	0.39 (0.489)	0.55	0.58
Republican†	0.328 (0.47)	0.354 (0.479)	0.62	0.53
Independent†	0.271 (0.445)	0.209 (0.407)	-1.66	0.10
High school or lower education†	0.321 (0.468)	0.268 (0.444)	-1.32	0.19
Some college†	0.313 (0.465)	0.37 (0.484)	1.37	0.17
Four-year degree†	0.229 (0.421)	0.248 (0.433)	0.51	0.61
Employed†	0.794 (0.405)	0.803 (0.398)	0.26	0.79

Table A4: Covariate balance for demographic covariates collected in survey. Note that indicators with a † were collected after the administration of treatment.

	Pre- and post-treatment belief measurement			
	(1)	(2)	(3)	(4)
Female	0.041	0.040	0.028	0.026
	(0.044)	(0.044)	(0.045)	(0.046)
Age	-0.001		-0.001	
	(0.001)		(0.001)	
Age 30-40		0.027		0.041
		(0.077)		(0.078)
Age 40-50		-0.044		-0.037
		(0.080)		(0.080)
Age 50-60		-0.027		-0.021
		(0.077)		(0.078)
Over age 60		-0.032		-0.033
		(0.071)		(0.072)
Republican			0.008	0.005
			(0.052)	(0.052)
Independent			-0.079	-0.084
			(0.057)	(0.058)
Some college			0.088	0.090
			(0.051)	(0.051)
Four year degree or more			0.069	0.069
			(0.057)	(0.057)
Employed			0.006	0.003
			(0.057)	(0.058)
(Intercept)	0.525	0.489	0.501	0.460
	(0.072)	(0.061)	(0.103)	(0.088)
Num. Obs.	516	516	516	516
Fstatistic	0.710	0.405	1.044	0.820
p-value	0.492	0.845	0.399	0.609

Table A5: Omnibus balalance test for demographic covariates collected in the survey. Columns 1-2 include only covariates measured pre-treatment; columns 3-4 include all demographic covariates. The F-statistic and its associated p-value serve as an test of balance across all covariates included in the model.

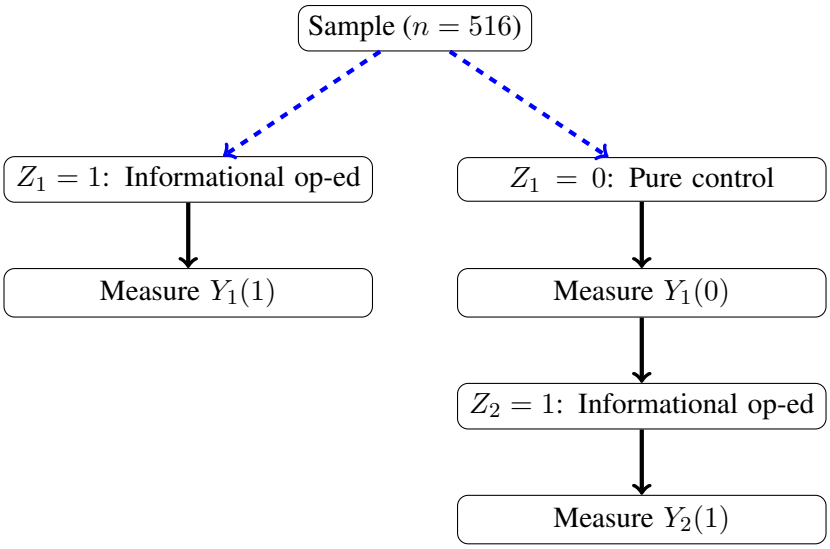

Figure A5: All blue dashed lines indicate random assignment. The 516 subjects are randomly assigned to each arm of the design using simple random assignment with probability $1/2$.

1. **Consent**: All participants included the survey provided informed consented to participate in the survey on the first page of the online survey instrument. Consistent with the APSA guidance, the consent form contained the researcher (PI) name and affiliation; an explanation that the survey was measuring "beliefs about climate change and preferred climate policies" and that participants would be asked to "read an opinion article about climate change;" discussion of the benefits (knowledge) and harms (potential for a breach of confidentiality) of participation; and notice that participants could withdraw at any point or skip any question.

2. **Deception**: The only deception involved the lack of revelation of the purpose of the op-ed. I remedy this through a debriefing script at the end of the survey (quoted below):

> Thank you for your participation in this research study. For this study, it was important that information was withheld from you. Now that your participation is complete, you will be given more information about this.
>
> **What you should know about the study**: Specifically, this study sought to measure how exposure to the op-ed changes beliefs about climate change and attitudes toward climate policy. Revealing this goal at the beginning of the survey may have changed how some participants respond to the op-ed.
>
> Because you were not fully informed about the study during the consent process, you may choose to have your data not used for the research analysis. Please indicate below if you give permission to have your data included in the research analysis:
>
> - I give permission for my data to be included in the study.
> - I DO NOT give permission for my data to be included in the study."

Withdrawal rates were low on average (3.7%) and did not detectably vary across the two treatment arms.[5]

3. **Confidentiality**: The survey was designed such that the personally identifying information was visible only to the survey firm (not the researcher), the survey instrument collected no personally identifying information (PII), and the survey responses were visible only to the researcher (not the survey firm). As a result, there was not a way to link the PII to responses, thereby preserving the confidentiality of responses.

4. **Harm**: The researcher did not conjecture any substantial harms to subjects from participation in the survey. As disclosed in the consent form, the most substantial potential for harm would be a breakdown in efforts to separate PII from survey responses. The research team is not aware of any such breach.

5. **Participant compensation**: All participants that offered informed consent ere compensated through the online panel through which they were recruited. Each participant received the same compensation, regardless of the number of questions answered. (The survey firm did not disclose the exact amount of compensation to the research team, though it was presumably less than $2.65 per respondent, the per-respondent cost paid by the researcher.)

6. **Participant characteristics**: The sampling frame sought a representative sample of US adults. The survey firm used gender and age to aid in assuring such a sample. Table A4 reports the treatment-condition means across these characteristisics, partisanship, and education. These descriptive data suggest that the share of vulnerable or marginalized groups is not higher than in the population as a whole.

A8.1.3 Results

Figure A6 reports standardized ATE estimates for each component of the index as well as the pre-specified primary outcome measure, the Z-score index of these four measures. For each outcome there are two estimates of the ATE. One comes from the experimental design which compares the pre-treatment beliefs in the pre-post group (e.g., $Y_i(0)$) to the post-treatment beliefs of the post-only group (e.g., $Y_i(1)$). The second comes from the within-subjects estimates that compares pre- to post-treatment measures of beliefs for all individuals in the pre-post group. Two findings are immediately evident: on average, the climate change op-ed increases beliefs about the severity of climate change, the role of human activity as a cause of climate change, prioritization of new policies on climate change, and the priority of investing in renewable sources of energy. This is evident from the positive point estimates. Second, the estimates from the within-subjects design is substantially more precise than those from the experimental design, as shown by the narrower confidence intervals.

Table A6 reports the estimates that are used for the decision rule in a regression table. The quantities used in the decision rule are shaded in blue. All quantities are estimated using OLS. Columns 1-2 report the two estimates of the ATE of the informational op-ed compared to a pure control group (no information). These estimates are plotted for the "Index" outcome in Figure A6. Column 3 reports the bias induced by the

[5]Note that if attrition were asymmetric across treatment arms, it would likely bias the decision rules *against* the within-subjects design because bias from attrition would be conflated with bias from the repeated treatments/questions.

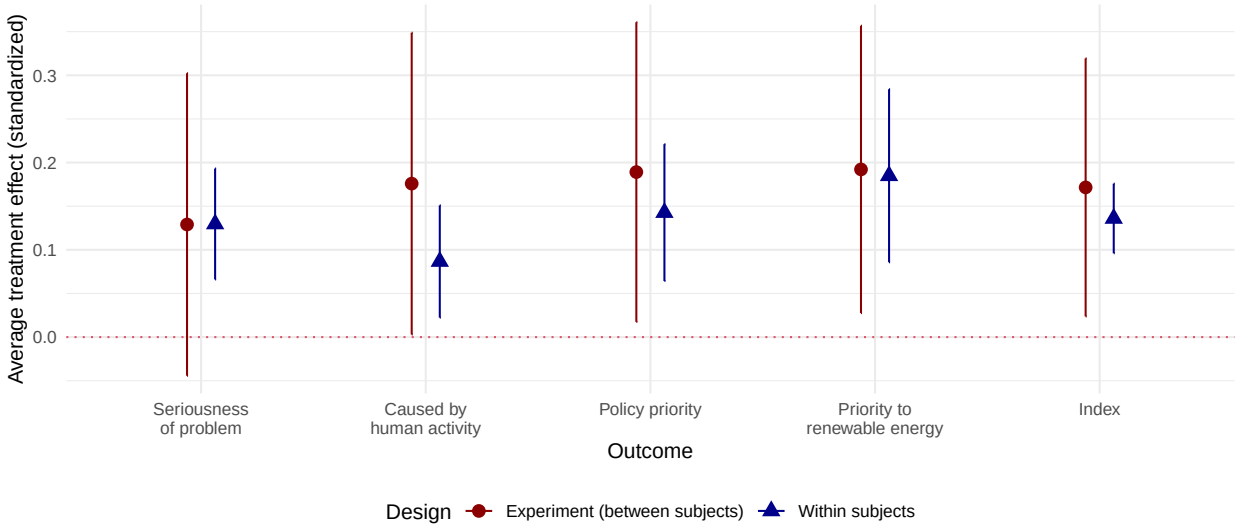

Figure A6: Estimated average treatment effects (ATEs) on each standardized component of the index as well as the Z-score index. The segments depict (unadjusted) 95% confidence intervals that were constructed on heteroskedasticity-robust standard errors.

sequential treatments in the within-subjects design.

The main paper shows that both decision rules select the within-subjects design for the pre-specified index outcome. Figure A7 provides an exploratory application of the decision rules to each component outcome. For the outcomes for which there is not evidence of bias (seriousness of the problem, policy priority, and priority afforded to renewable energy), both decision rules unsurprisingly select the within subjects design. In contrast, the ATE estimates from the experimental and within subjects designs depart for the belief that climate change is caused by human activity. Specifically, the ATE from the within subjects design is about half of the ATE from the experimental design. In this case, decision rule #1 (sign) selects the experimental design whereas decision rule #2 (point estimate) selects the within subjects design.

A8.2 Application #2: Welfare vs Support to the Poor

A8.3 Design

Clifford, Sheagley, and Piston (2021) (Study 1) conduct a question wording/framing experiment that contains a design parallel to Hybrid Design #2. The design of the survey and the responses that I analyze are depicted on the right of the tree in Figure A9. From inspection of the data and replication materials, it appears that the authors used simple random assignment with probability $1/2$ at each node.

A8.4 Results

Table A7 reports the estimates that are used for the decision rule in a regression table. All quantities used in the decision rule are shaded in blue. All estimates are estimated using OLS. Columns 1-2 report the two estimates of the ATE of the welfare question compared to the question about support for the poor. Columns

	(1)	(2)	(3)
	Exp. ATE $Y_{i1}(Z)$	Within ATE $Y_{it}(1) - Y_{i,\neg t}(0)$	Bias $Y_{it}(1)$
Intercept	3.620	0.136	3.890
	(0.172)	(0.020)	(0.055)
Informational op-ed	0.172		-0.036
	(0.075)		(0.075)
Num. Obs.	516	254	516
Comparison is	Experimental	Observational	Experimental
Naïve 95% CI:	[0.023, 0.320]	[0.096, 0.176]	
Bonferroni 95% CI:	[0.002, 0.341]	[0.090, 0.182]	

Table A6: Quantities that enter decision rule are highlighted in blue. All regressions are estimated using OLS. Heteroskedasticity robust standard errors in parentheses.

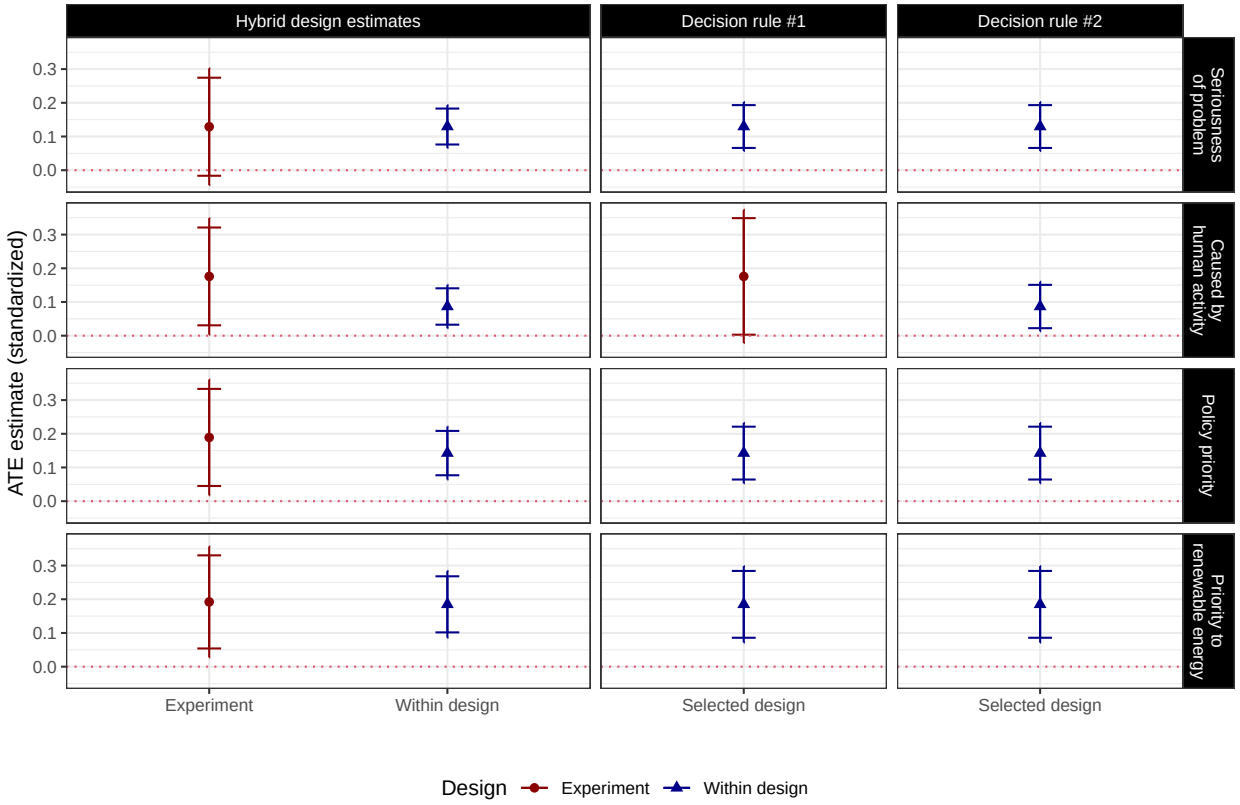

Figure A7: Exploratory application of the decision rules to each of the component outcome measures. Estimates of the ATE from both designs are in the left panel. The center and right panels show the application of each decision rule. Vertical segments denote 95% confidence intervals. The horizontal marks show 90% confidence intervals. The confidence interval length is greater in the right panels due to the Bonferroni correction invoked in the hybrid design.

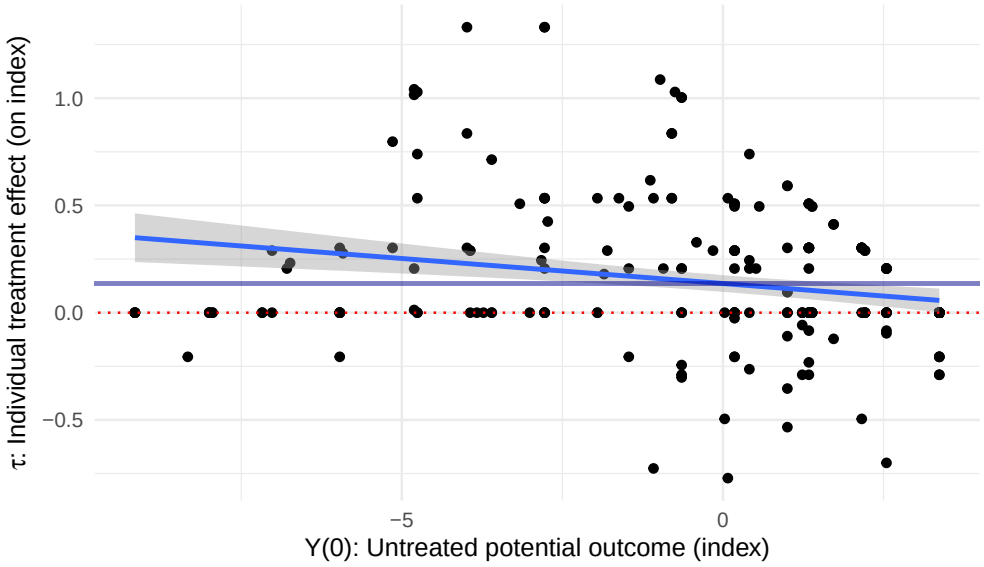

Figure A8: This figure plots the relationship between untreated potential outcomes ($Y_i(0)$) and individual treatment effects (τ_i) from the subjects assigned to the within-subjects design. The estimated slope of the blue line is -0.022 (95% CI: [-0.035,-0.012]), documenting a negative covariance between these parameters that is consistent with Bayesian updating. The purple line depicts the estimated ATE.

3-4 report the bias induced by the sequential treatments in the within-subjects design. There is very limited evidence that the within-subjects designs induces bias. The bottom panel denotes whether each comparison is experimental versus observational. It then compares the naive 95% CIs to the Bonferroni-corrected 95% CIs on the ATE estimate that must be used when the decision rule is employed to select a design.

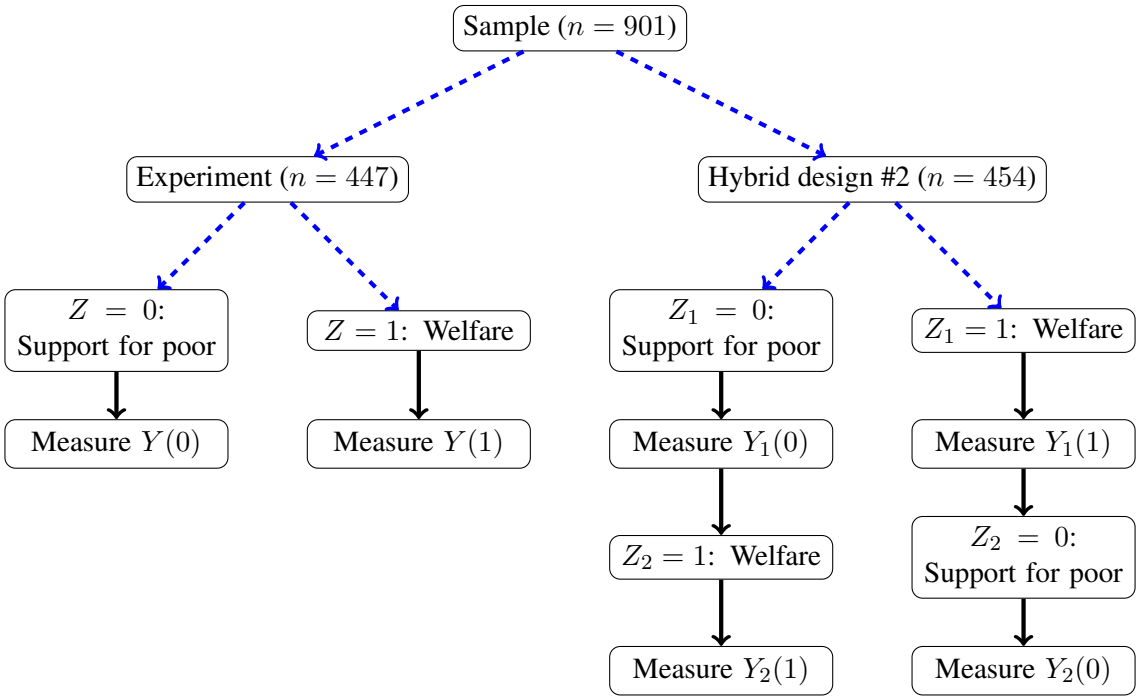

Figure A9: All blue dashed lines indicate random assignment. I analyze only the hybrid design on the right hand side of the tree.

	(1)	(2)	(3)	(4)
	Exp. ATE $Y_{i1}(Z)$	Within ATE $Y_{it}(1) - Y_{i,\neg t}(0)$	Bias 1 $Y_{it}(1)$	Bias 2 $Y_{it}(0)$
Intercept	2.500	-0.278	2.195	2.500
	(0.044)	(0.033)	(0.048)	(0.044)
Welfare question	-0.289		0.017	-0.035
	(0.070)		(0.072)	(0.064)
Num. Obs.	453	453	453	454
Comparison is	Experimental	Observational	Experimental	Experimental
Naïve 95% CI:	[-0.427 -0.151]	[-0.342 -0.213]		
Bonferroni 95% CI:	[-0.446, -0.131]	[-0.351, -0.204]		

Table A7: Quantities that enter decision rule are highlighted in blue. All regressions are estimated using OLS. Heteroskedasticity robust standard errors in parentheses.

Supplementary Appendix: References

Arnesen, Sveinung, Dominik Duell, and Mikael Poul Johannesson. 2019. "Do citizens make inferences from political candidate characteristics when aiming for substantive representation?" *Electoral Studies* 57: 46–60.

Atkeson, Lonna Rae, and Brian Hamel. 2020. "Fit for the Job: Candidate Qualifications and Vote Choice in Low Information Elections." *Political Behavior* 42: 59–82.

Banks, Antoine J., and Heather M. Hicks. 2016. "Fear and Implicit Racism: Whites' Support for Voter ID Laws." *Political Psychology* 37 (5): 641–658.

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.

Blackman, Alexandra Domike, and Marlette Jackson. 2021. "Gender Stereotypes, Political Leadership, and Voting Behavior in Tunisia." *Political Behavior* 43: 1037–1066.

Boas, Taylor C., F. Daniel Hidalgo, and Marcus André Melo. 2019. "Norms versus Action: Why Voters Fail to Sanction Malfeasance in Brazil." *American Journal of Political Science* 63 (2): 385–400.

Carey, John, Katherine Clayton, Gretchen Helmke, Brendan Nyhan, Mitchell Sanders, and Susan Stokes. 2022. "Who will defend democracy? Evaluating tradeoffs in candidate support among partisan donors and voters." *Journal of Elections, Public Opinion & Parties* 32 (1): 230–245.

Christenson, Dino P., and Douglas L. Kriner. 2019. "Constitutional Qualms or Politics as Usual? The Factors Shaping Public Support for Unilateral Action." *American Journal of Political Science* 61 (2): 335–349.

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.

DeScioli, Peter, Alex Shaw, and Andrew W. Delton. 2018. "Share the Wealth: Redistribution Can Increase Economic Efficiency." *Political Behavior 2018* 40: 279–300.

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.

Kirkland, Patricia, and Alexander Coppock. 2018. "Candidate Choice Without Party Labels: New Insights from Conjoint Survey Experiments." *Political Behavior* 40: 571–591.

Pérez, Efrén, and Margit Tavits. 2019. "Language Influences Public Attitudes toward Gender Equity." *The Journal of Politics* 81 (1): 81–93.

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).

Senninger, Roman, and Daniel Bischof. 2023. "Do Voters Want Domestic Politicians to Scrutinize the European Union?" *Political Science Research and Methods* 11 (2): 410–418.