Supplementary Materials for

"Does Conjoint Analysis Mitigate Social Desirability Bias?"

Yusaku Horiuchi Zachary Markovich Teppei Yamamoto

A Limitations of the Existing Literature

There exists a well-developed literature examining whether factorial survey designs (Wallander 2009) mitigate SDB through mechanisms similar to the ones discussed above (e.g. Atzmüller and Steiner 2010). Although conjoint analysis is a specific type of factorial survey, we argue that the existing studies on this topic do not provide sufficient empirical evidence to confirm the SDB-mitigating effect of a fully randomized conjoint design. This is due to shortcomings of these previous studies and specific characteristics of the fully randomized conjoint design.

First, previous research fails to distinguish a reduction in SDB from *design effects*, or discrepancies in respondents' observed preferences due to differences in survey designs unrelated to social desirability. Most empirical approaches to evaluating the SDB-mitigating effects of factorial designs rely on the comparison between preferences elicited through a factorial design with an alternative question format (e.g., Auspurg et al. 2014). For example, in some studies, respondents express weaker preferences for gender pay equality and more tolerance for unethical business practices in factorial experiments than in direct questioning, leading researchers to conclude that conjoint analysis has uncovered less biased attitudes (Auspurg, Hinz and Sauer 2017; Jasso and Webster Jr 1999). However, a major limitation of such analyses is their inability to distinguish a reduction in SDB from other effects that question formatting may have on responses.

Second, existing research on SDB focuses on vignette-based factorial surveys rather than the simplified conjoint tables typically employed in political science. Recent evidence (Jenke et al. 2020) suggests that tabular conjoint surveys may *not* mitigate SDB as effectively as vignettes since respondents are adept at focusing on the small number of attributes most relevant to them in tabular surveys. Narrative vignette designs (e.g., Rossi and Nock 1982) may have an advantage over tabular designs by making the sensitive attribute even less noticeable. However, to the best of our knowledge, no existing study directly examines the

effectiveness (or ineffectiveness) of reducing SDB when respondents are asked to complete fully randomized conjoint tasks in the tabular format.

B More on Topic Selection

In an early stage of our project, we contemplated using several alternative attributes (and their corresponding choice settings) that also seemed plausible, such as the race (Abrajano, Elmendorf and Quinn 2018) and gender (Teele, Kalla and Rosenbluth 2018) of political candidates. However, after thinking systematically about the criteria for a substantive topic to be used in the design, and considering the recent empirical literature on race and gender, we concluded that neither of these topics would be suitable for our study.

For race, growing evidence suggests that the norm of racial equality (Mendelberg 2001) has been rapidly eroding among American conservatives in recent years (Tesler 2016; Valentino, Neuner and Vandenbroek 2018). Hence, conservatives may not hesitate to express their true preferences (e.g., in favor of white political candidates) regardless of survey format. On the other hand, many liberals consider racial diversity in political elites to be of primary, rather than secondary, representational concern when making their vote choices. Therefore, unlike eco-conscious respondents' decisions about purchasing athletic shoes, liberals' vote choices will generally match the direction of their true — and socially desirable — racial preference (e.g., in favor of non-white candidates). After excluding these two groups of respondents who are not susceptible to SDB, we are likely to be left with an unacceptably small sample for our analysis. Compounding this difficulty, separating liberal respondents who genuinely prefer non-white candidates from those who prefer white candidates but avoid choosing them because of SDB would be difficult using observable variables. This is because their self-reported racial attitudes will be observationally equivalent; for example, their responses to a standard racial resentment battery (Kinder 2013) will be difficult to distinguish from each other, both reporting low levels of racial resentment.

Similar concerns about obtaining a sufficiently large number of effective responses apply

to experiments on the gender of a candidate. Moreover, the growing body of mixed evidence on the effects of gender on voter preferences using fully randomized conjoint designs (e.g., Schwarz and Coppock 2020) suggests that there is substantial heterogeneity in voters' preferences about candidate gender conditional on a multitude of respondent characteristics. This complexity would add to the difficulty of pinning down a specific subgroup of respondents for our study. We note, however, that the social sensitiveness of the sexual harassment attribute in our Study 2 emanates from the norms of gender equality. Therefore, our results from Study 2 may well be generalizable to experiments on a candidate's gender.

C More Details and Additional Results for Study 1

C.1 Block Randomization

We implemented a block randomization strategy to eliminate potential imbalances in observed covariates and to improve efficiency in estimation. Specifically, we stratified respondents into blocks using the covariates, such that respondents are identical in terms of those variables within each block. We then completely randomized them into the four design conditions with equal probability within each block, so that the resulting treatment groups are nearly perfectly balanced with respect to those covariates.

In Wave 1, we measured many demographic variables of respondents and their political attitudes. We constructed blocks based off of this information. Specifically, we blocked on (1) age, (2) race, (3) partisanship, (4) environmental attitudes, and (5) SDB proneness. Our battery of general SDB questions consisted of eight items from the impression management scale in the Balanced Inventory of Desirable Responding Short Form (BIDR-16, Hart et al. 2015). In addition, we included one item from the longer BIDR-40 (Paulhus and Reid 1991) that asked about a subject's propensity to litter, which we saw as being directly relevant to a respondent's likelihood of stating a dishonest preference for an eco-friendly product. The subgroup of respondents we were primarily interested in are those who are both SDB

prone and not anti-environment. We chose the other three blocking covariates (i.e., age, race, and partisanship) because substantively we believe that they are likely to be correlated with respondents' attitudes about the environment.

When creating blocks, we coarsened age categories so that it represented whether or not a respondent was over 35 years old, and race categories so that it represented whether or not a respondent was white. We blocked on partisanship based on whether a respondent was a Democrat, a Republican, or something else. We counted respondents who indicated that they leaned towards one party or the other as members of that party and also grouped respondents who identified as independents or as members of a third party together in the third category. We defined a respondent as holding an anti-environment attitude if they chose the most anti-environmental option in any of the five eco-friendliness questions that we asked. We first dichotomized each of the eight seven-point-scaled SDB items such that the most, the second most, and the third most socially desirable options are coded as a "SDB-prone" response. We categorized respondents as SDB prone if they registered an SDB prone response on four or more of the eight SDB questions (see Section 5.1 for additional information).

Table C.1 shows the numbers of respondents in each of the 47 uniquely defined blocks based on the five blocking variables for Wave 1 (i.e. before attrition) and Wave 2 (after attrition). Note that Block 11 contains both age groups, because the older group (36 years old or older, non-white, independent, anti-environment, not SDB-prone) turns out to contain only two observations based on the Wave 1 data. We assign the design conditions by complete randomization within each of these 47 blocks as respondents answered the Wave 2 questions. That is, these five covariates are nearly perfectly balanced within the Wave 2 sample.

C.2 Estimation Methodology

The estimates reported in Section 6.1 are obtained via a variant of the least squares estimator proposed by Hainmueller, Hopkins and Yamamoto (2014), which incorporates the design

Table C.1: Block Randomization

Block ID	Age	Race	Partisanship	Anti Environment	SDB Prone	N, Wave 1	N, Wave 2
1	0	0	1	0	1	172	152
2	0	0	2	0	1	31	26
3	0	0	3	0	1	40	36
4	0	1	1	0	1	298	262
5	0	1	2	0	1	185	159
6	0	1	3	0	1	64	55
7	1	0	1	0	1	114	103
8	1	0	2	0	1	29	25
9	1	0	3	0	1	20	19
10	1	1	1	0	1	345	322
11	1	1	2	0	1	232	209
12	1	1	3	0	1	82	76
13	0	0	1	0	0	140	122
14	0	0	2	0	0	28	22
15	0	0	3	0	0	17	16
16	0	1	1	0	0	241	207
17	0	1	2	0	0	113	92
18	0	1	3	0	0	38	33
19	1	0	1	0	0	61	55
20	1	0	2	0	0	17	15
21	1	0	3	0	0	6	6
22	1	1	1	0	0	174	164
23	1	1	2	0	0	85	77
24	1	1	3	0	0	26	19
25	0	0	1	1	1	38	33
26	0	0	2	1	1	18	14
27	0	0	3	1	1	10	9
28	0	1	1	1	1	66	60
29	0	1	2	1	1	66	54
30	0	1	3	1	1	16	12
31	1	0	1	1	1	27	24
32	1	0	2	1	1	10	10
33	1	0	3	1	1	9	9
34	1	1	1	1	1	77	72
35	1	1	2	1	1	87	83
36	1	1	3	1	1	25	23
37	0	0	1	1	0	64	55
38	0	0	2	1	0	24	22
39	0/1	0	3	1	0	13	12
40	0	1	1	1	0	80	63
41	0	1	2	1	0	76	64
42	0	1	3	1	0	15	15
43	1	0	1	1	0	15	12
44	1	0	2	1	0	10	8
45	1	1	1	1	0	83	73
46	1	1	2	1	0	66	59
47	1	1	3	1	0	18	16

Note: Block 39 contains two observations, which differ only in terms of Age (1 rather than 0). We merged a block with less than four observations so that each block has at least as many observations as the number of the design conditions (four).

conditions as well as the block randomization. Specifically, we first recode the treatment and control attribute dummy variables into an "always varying attribute" dummy (A1) and a "not always varying attribute" dummy (A0) based on the design conditions. That is, A1 and A0 are respectively equal to the treatment and control attribute dummies in the Partial-Sensitive and Full-Sensitive conditions, and vice versa in the Partial-Control and Full-Control conditions. Then, we regress the observed seven-point outcome variable on A1, the partial condition dummy, the treatment condition dummy, all possible interaction terms for the above, A0, a set of dummies for the filler attributes, and a set of dummies for block randomization. The AMCE estimates can then be obtained as corresponding linear combinations of least squares coefficients on A1 and its interactions with the design dummies.

C.3 Conjoint Attributes

Table C.2 shows the full list of attributes used in Study 1.

C.4 Additional Results

As noted in Section 6.1, the statistically significant SDB-mitigating effect (Figure 2) disappears when we conduct the same analysis on the SDB-proof respondents. As presented in Figure C.1, the AMCE for the placebo attribute is 1.27 ($[1.13, 1.41]$) and 0.40 ($[0.32, 0.48]$) under the partial and fully randomized conditions, respectively, which amounts to the difference of 0.87 between the two design conditions. The corresponding estimates for the sensitive attribute are 1.47 ($[1.29, 1.65]$) and 0.48 ($[0.39, 0.56]$), resulting in the difference of 0.99. The difference between these two differences (i.e., $(1.47 - 0.48) - (1.27 - 0.40) = 0.13$) is statistically insignificant with the 95% confidence interval of $[-0.13, 0.39]$.

One possible threat to our inference is differential respondent fatigue between the sensitive and placebo conditions: the partially randomized designs may induce respondents to satisfice more than the fully randomized designs because the tasks might feel more repetitive in the partially randomized conditions. To test this possibility, we investigate whether the difference-in-differences in the AMCEs across the four design conditions grew over the course

Attribute	Levels
Eco-Friendly Materials (**sensitive**)	100% Eco-Friendly Materials Used
	No Eco-Friendly Materials Used
Gel Cushioning (**placebo**)	Has Gel Cushioning
	No Gel Cushioning
Brand	Nike, Adidas, Vans, Puma, Under Armour, Reebok
Model Year	2019, 2018, 2017, 2016
Ave. Customer Review	5 out of 5, 4.5 out of 5, 4 out of 5, 3.5 out of 5
Price	$110, $88, $64, $43
Color	Gray, White, Navy, Red
Shipping	Free Standard Shipping, Free Expedited Shipping,
	Additional Shipping Charges Apply
Weight	5 oz., 7 oz., 9 oz., 11 oz.
Best Seller	#1 in Athletic Shoes, #5 in Athletic Shoes,
	#12 in Athletic Shoes, #55 in Athletic Shoes,
	#100 in Athletic Shoes, #250 in Athletic Shoes

Table C.2: List of Attributes for Study 1

of the twenty tasks each respondent completed. The test for a linear trend in the difference-in-differences estimate fails to reject the null of no effect ($p < 0.25$) and these results are similar when the task number is coarsened into a four level categorical variable.

Another potential problem is that the direct questions about interest in protecting the environment, which we used to construct the SDB-prone subgroup of respondents, may not be sufficient as a measure of SDB-proneness. As a validity check, we examine the AMCE for the eco-friendly materials among respondents who express "anti-environment" responses. When we use a strict criterion (i.e., only respondents expressing the strongest preference against environmental protection in at least three of our five direct question items), the estimated AMCE of the eco-friendly material is negative and statistically indistinguishable from zero (-0.20). With a less strict definition (i.e., expressing the strongest preference

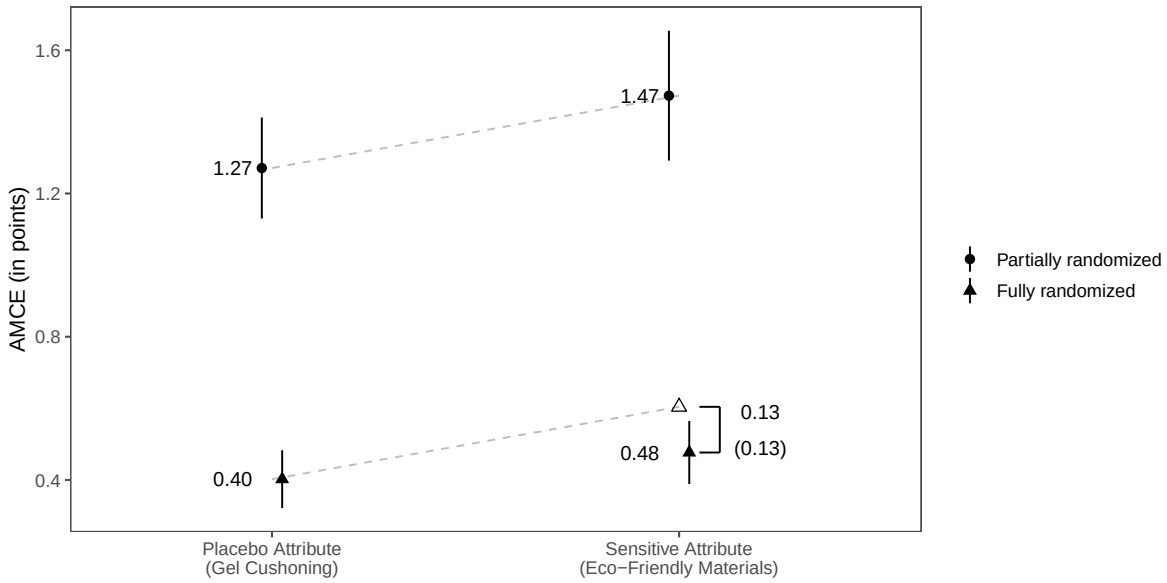

Figure C.1: Average Marginal Component Effects (AMCEs) for the Sensitive and Placebo Attributes under the Partial and Fully Randomization Designs in Study 1. *See the caption for Figure 2 for the explanations of the graph elements. The outcome variable is a 7-point Likert Scale measure of preference for hypothetical athletic shoes (least likely to purchase* $= 1$*; most likely to purchase* $= 7$*). This figure uses all respondents identified as not sensitive to SDB.*

against environmental protection in at least two of our five direct question items), it becomes positive (0.30) but still statistically insignificant and substantially smaller than the estimate reported in Figure 2.

D More Details and Additional Results for Study 2

D.1 The SDB-Priming Treatment

Before the conjoint tasks, we showed a random half of respondents (i.e., the treatment, *sensitive* group) a paragraph stating that we might contact them again for a follow-up survey with an invitation to complete a face-to-face interview (Figure 1). The other half (i.e., the *control* group) were given no such treatment and proceeded directly to the conjoint tasks. To increase this stimulus's strength, we also asked a question, "Would you be interested in

this face-to-face interview?" Respondents answering "Yes" were directed to the next screen, saying, "Thank you for expressing your interest in the face-to-face interview." Respondents answering "No" were directed to a different screen, saying, "Thank you for your response. If you change your mind, please note that your interest in the face-to-face interview in the comment box at the end of this survey." By repeatedly emphasizing the possibility of a *face-to-face interview*, we induced additional social pressure on respondents in the treatment group, expecting that their evaluation of hypothetical candidates would be more biased toward the direction of social desirability (i.e., lower ratings for candidates with a sexual harassment scandal).

D.2 Conjoint Attributes

Table D.1 shows the full list of attributes used in Study 2.

Attribute	Levels
Scandal (**sensitive**)	None, Sexual Harassment
Previous Profession	Community Organizer, Business Executive, Military
Past Political Experience	3 years, 5 Years, 11 Years
Undergraduate Degree	Community College, State University, Ivy League Degree
Age	50, 55, 62, 65, 67
Race	White, Black, Hispanic, Asian American
Gender	Man, Woman
Residency	Lives Inside Your Congressional District, Lives Outside Your Congressional District

Table D.1: List of Attributes for Study 2

D.3 Selection of Survey Platform

We switched away from recruiting respondents from MTurk, given the concern that the quality of MTurk respondents has worsened recently (Kennedy et al. 2020). We explored

multiple platforms in our pre-tests, including Lucid Theorem, which seemed to have growing popularity. However, only a low percentage of Lucid respondents passed our simple attention check question (also see Peyton, Huber and Coppock 2020). Prolific claims that they only recruit high-quality MTurk workers. Indeed, our data show high percentages of clearing the attention check and correctly answering the factual manipulation check. There is also some corroborating evidence in the recent literature (Adams, Li and Liu 2020; Palan and Schitter 2018).

D.4 Construction of the SDB-prone Subgroup

For our analysis in the main paper (Figure 3), we used the causal forest to identify a subset of respondents for which the prime was successful in inducing more socially desirable responding (Wager and Athey 2018). The causal forest is well suited for our purposes for several reasons. First, as noted in the main paper, it easily generates out-of-bag predictions of the effect of the prime on each respondent. These predictions are made without using the respondent for which the prediction is being made, so the main analysis testing for the SDB-reducing effect of conjoint analysis can be conditioned on these estimates without biasing the results (Athey and Imbens 2016). Second, tree-based methods, of which the causal forest algorithm is an example, are well known for their good performance on datasets where there is little theoretical basis to parametrically model the relationship of interest but only a moderate number of observations (Montgomery and Olivella 2018). Additionally, causal forests produce consistent estimates of treatment effect heterogeneity so this algorithm will asymptotically identify the true set of SDB prone respondents (Athey, Tibshirani and Wager 2019). Finally, the causal forest algorithm is implemented in the well maintained and highly optimized `grf` package which we used for our analysis (Tibshirani et al. 2020).

We made several changes to the default model parameters used by the implementation of the causal forest in the `grf` package. The motivation for these changes stems from the fact that we are interested in using the estimates of treatment effect heterogeneity as a

conditioning variable for our main analysis rather than the primary quantity of interest themselves, which is the purpose the default parameters are optimized for. In particular, this means that we are not concerned with bias in the estimates of treatment effect heterogeneity and do not require variance estimates to be associated with the estimates for the effect of the prime on each respondent. Consequently, we are willing to tolerate a small amount of bias in our estimates and disable variance estimation in order to improve the overall accuracy of the estimates of treatment effect heterogeneity. Specifically, we made the following changes to the algorithm's default settings:

- Disabling the "honesty" option – the default settings of the causal forest algorithm generate forests that are "honest" in the sense of Athey and Imbens (2016). Honest forests are beneficial in many settings because they provide unbiased estimates of treatment effect heterogeneity, but come with the drawback of increasing the variance of those estimates. In our empirical strategy, the treatment effect heterogeneity itself is not the ultimate quantity of interest. Rather, the causal forest is only used to construct an SDB-prone subset of respondents in the first stage of our analysis. We therefore do not require unbiasedness in the causal forest and so we disabled honesty to reduce the total mean squared error of our estimates. This represents an instance of the bias-variance trade-off where tolerating a small amount of bias in the estimates of treatment effect heterogeneity can dramatically reduce their variance and improve overall performance. This choice is also in line with recommendations made by the `grf` package developers who recommend disabling honestly when working with relatively small datasets.

- Eliminating the sample split used for variance estimation – the causal forest algorithm generates variance estimates for its estimates of treatment effect heterogeneity using sample splitting. Specifically, for each tree grown, it uses a subset of the data for estimating the treatment effect heterogeneity itself and the remainder for estimating

the variance of those estimates. Since we do not use these variance estimates in our analysis, the statistical efficiency of the treatment effect estimates can be improved by disabling variance estimation and using the full dataset for estimating treatment effect heterogeneity. This choice is again in line with the recommendations of the package developers who suggest this choice for improving model performance on small datasets.

- Increasing the fraction of the dataset used to grow each tree to .9 – the causal forest algorithm is based on the concept of subsampling. The "forest" is composed of many different causal trees trained on different random subsamples of the data and the size of these subsamples is controlled by one of the algorithm parameters. Variance estimates are then generated by further subsampling from the remaining data. Larger subsamples will result in more precise estimates, but the `grf` package caps the size of these subsamples at .5 by default to ensure that there is sufficient data available for variance estimation. Since our modeling approach makes no use of these variance estimates, we increased the fraction of the dataset used for growing each tree to .9.

- Increasing the number of trees to 20,000 – to eliminate noise introduced into our estimates by the subsampling and tree growing procedure used by the causal forest, we increased the number of trees used for the forest to 20,000. This is also in line with a standard recommendation in machine learning literature to choose the largest computationally feasible number of trees for a random forest (Probst and Boulesteix 2017).

Our measure of SDB-proneness is the predicted treatment effect of the prime on the AMCE for the scandal attribute in the partially randomized condition. We could consider other measures. For example, our survey includes a post-treatment SDB-proneness battery similar to those used in the first wave of Study 1. However, we consider that our chosen measure is the most relevant for several reasons. First, the partially randomized design blocks the SDB mitigating effect of conjoint analysis, so the prime's SDB-increasing effect is most

clearly displayed in this design condition. Second, the AMCE for the scandal attribute directly reflects how socially sensitive a respondent finds the conjoint tasks themselves. At the same time, other post-treatment measures of SDB could indicate SDB proneness more generally and only indirectly reflect respondents' perceived sensitivity of conjoint tasks. Third, importantly, our prime specifically referred to the conjoint evaluations of candidate profiles as the subject of the face-to-face interview the respondents were potentially invited to (see Figure 1). Finally, the effect of the prime should be more clearly expressed during the conjoint tasks that immediately followed the prime.

D.5 Estimation Methodology

Our estimation procedure for the SDB reduction is similar to the approach used in Study 1 (see Section C.2). The "always varying attribute" (A0) is whether a candidate is involved in a sexual harassment scandal. There is no placebo attribute in Study 2. As in Study 1, we regress the observed seven-point outcome variable on A1, the partial condition dummy, the treatment condition dummy, all possible interaction terms for the above, and a set of dummies for all the other attributes. We did not administer a block randomization in Study 2. The quantity of interest is the coefficient on the triple interaction term, as in Study 1.

D.6 Robustness to Alternate Model Parameters

While the default settings for the remaining model parameters are sensible and appear to perform well in our setting, we also explored robustness of our results to changes in these parameters. Specifically, we explored how the results varied with changes to the parameter `alpha` which controlled the maximum imbalance allowed for splits in the causal trees that compose the causal forest; `minimum node size`, which controls the minimum size allowed for nodes in the causal forest; and `mtry`, which controls how many variables are randomly

sampled to grow each tree.[1]

Figure D.1 presents estimates and standard errors from our specification in Figure 3 when the group of SDB prone respondents is identified using a causal forest with various values of `alpha`, `minimum node size`, and `mtry`. The estimates are overwhelmingly negative suggesting that our results are fairly robust to modest changes in these model parameters. Our main analysis used the default values of `alpha`, `minimum node size`, and `mtry` which are .05, 10, and 5 respectively. Higher values of `alpha` and `mtry` and lower values of `minimum node size` allow more flexibility in the modeling process, but also risk overfitting. Parameter values that allow more flexibility than the package defaults are associated with larger estimates for the reduction in SDB facilitated by conjoint analysis while those associated with less flexibility result in smaller ones. This suggests that SDB proneness is related to a complex interaction of the various demographic predictors and cannot be adequately modeled with less flexible approaches.

Although it is encouraging that our results are not overly dependent on a single choice of the causal forest model parameters, it is unsurprising that over-regularized model parameters will result in a poor performance. Indeed, for any possible pattern of treatment effect heterogeneity it will be possible to choose model parameters extreme enough that the causal forest is unable to model it. Consequently, it is better to evaluate results across a variety of reasonable model parameters rather than focusing on the existence of a subset of parameters which produce poor results.

One alternate approach to identifying model parameters is to use cross-validation to identify parameters which optimize out of sample predictive accuracy. Although commonly used for choosing model parameters for predictive tasks, such approaches are controversial when the target of inference is treatment effect hetereogeneity. While out of sample accu-

[1]The `grf` package also includes an parameter which penalizes imbalanced splits when growing each causal tree. This is labeled as an experimental feature and is disabled by default. We did not explore model performance when using this feature.

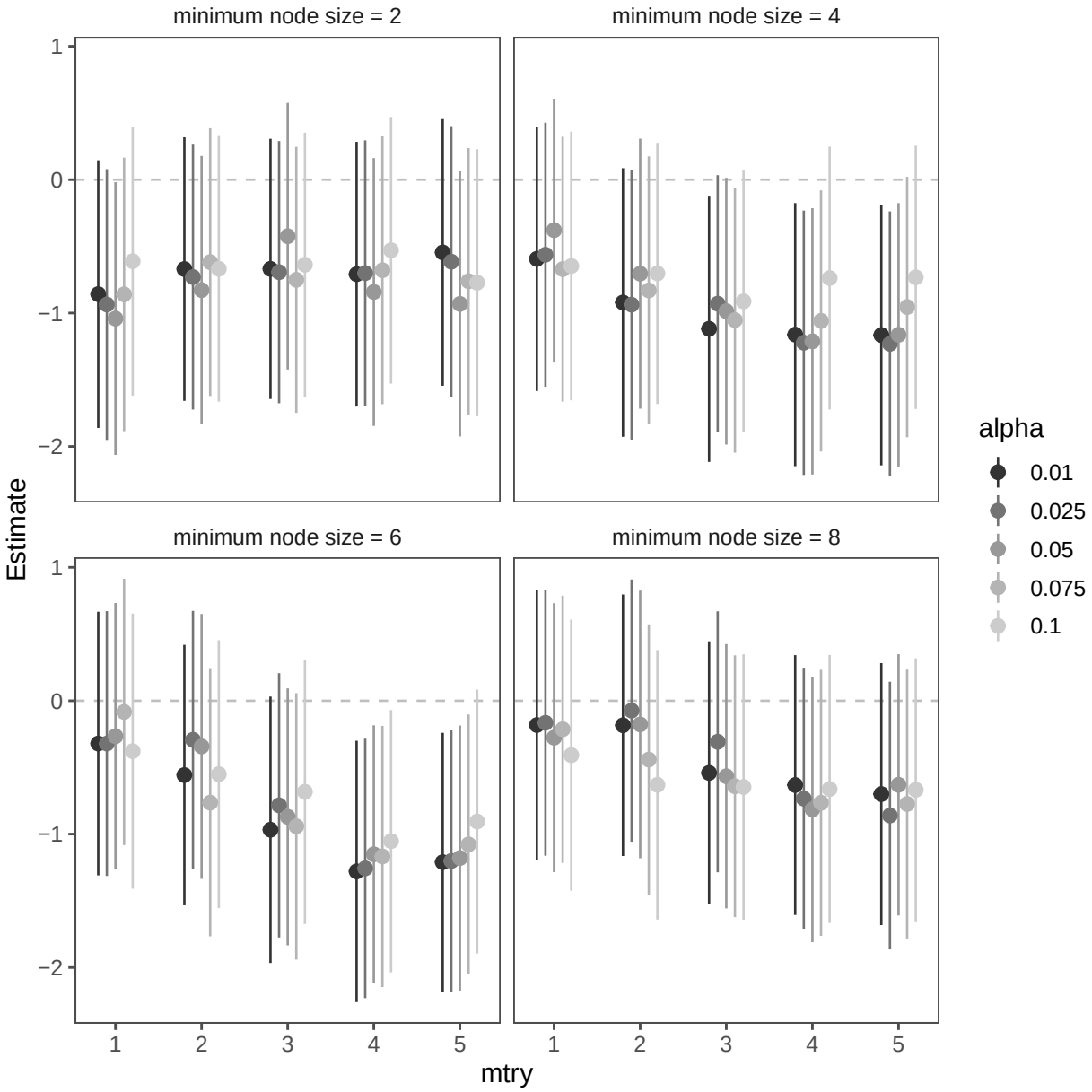

Figure D.1: Estimates for Reduction of SDB with Various Causal Forest Parameters

racy can be directly optimized for predictive tasks, such optimization is impossible in the context of causal inference as only one set of potential outcomes is ever directly observed. Consequently, there is no definite benchmark that can be optimized using cross-validation in this setting (Künzel, Walter and Sekhon 2019).

What is more, proposals that do exist for optimizing model parameters in the context of

treatment effect heterogeneity estimation are not well suited to our particular use case. Existent proposals for model tuning in the context of treatment effect heterogeneity estimation attempt to maximize the accuracy of predictions in the full sample (Nie and Wager 2020), but for our application, we are only interested in the accuracy of predictions about the subset of most SDB-prone respondents and do not care about the accuracy of predictions for the remainder of the sample. This distinction is particularly relevant for our application because our theoretical expectation is that there will be a small subset of respondents for which the prime had a negative effect on the AMCE associated with the scandal attribute in the constrained conjoint design, but that its effect will be null with no heterogeneity for majority of the sample. Consequently, cross validation approaches which focus on treatment effect heterogeneity in the full sample will over-regularize and choose parameter values that do not allow enough flexibility to accurately identify the subset of most SDB prone respondents.

Finally, cross-validation rotates which subset of the data is used as a training and test set and in the process chooses model parameters using the entirety of the dataset. However, our estimation strategy requires that the out-of-bag predictions of treatment effect heterogeneity for each point are independent of the outcome for that point, but cross-validation uses the full dataset to choose the model parameters, violating this assumption. While introducing a third data split to be used as a validation set could overcome this problem, such an approach is computationally infeasible due to the long training time needed to optimize even a single causal forest.

For these reasons, we have followed the recommendation of Künzel, Walter and Sekhon (2019) to present results for a wide range of sensible parameter values rather than focusing on a single set of parameters chosen via cross-validation.

D.7 Robustness to Alternative SDB-Prone Group Sizes

One choice in our analysis is the size of the SDB prone group to focus on. Estimates associated with a larger group will be more precise, but that larger group will also include

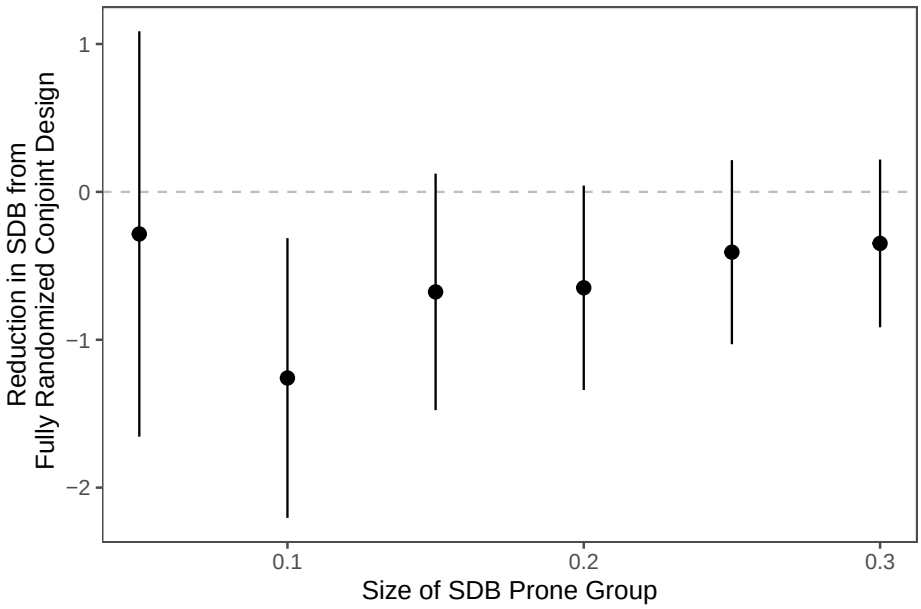

Figure D.2: Estimate For Reduction in SDB by Size of SDB Prone Group

respondents that were not as strongly influenced by the prime. We focus on the top .1 most SDB prone respondents in our main analysis as this choice seems to strike the best balance in terms of this trade off; however, Figure D.2 visualizes comparable estimates for a large number of groups between 0.05 and 0.3. Reassuringly, the magnitude and precision of the estimates vary as is expected. Estimates which are based on a smaller set of more SDB prone respondents suggest a greater reduction in SDB, but are also less precise.

D.8 Results Using Different Algorithms

An alternative methodology for identifying treatment effect heterogeneity is the set of meta-learners proposed by Künzel et al. (2019). Figure D.3 presents the results of this analysis for three meta-algorithms: the S-learner, T-learner, and X-learner. In all cases, a random forest is used as the base learner.[2] Because the choice of the size of the SDB prone group

[2]We use the implementation of these algorithms made available in the `causalToolbox` package. We make changes to the default model parameters in line with those described in Section D.4. Specifically, for all random forests used in the meta-algorithm we increase the fraction of data points sampled when

of respondents to focus on is somewhat arbitrary, we replicate our approach from Section D.7 and visualize estimates for many different sizes of the SDB prone group. The Künzel et al. (2019) meta-learners do not naturally produce out-of-bag predictions of treatment effect heterogeneity like the causal forest used in our main results does. Consequently, we proceed by randomly splitting respondents into ten folds and generate predictions of treatment effect heterogeneity for each fold using only respondents assigned to one of the other folds. This sample splitting procedure is less efficient than the out-of-bag predictions generated by the causal forest because it only uses nine tenths of the respondents when generating the prediction for each fold, while the out-of-bag predictions use all respondents except for the one that predictions are being made for.

While all three approaches generate point estimates consistent with the fully random conjoint design reducing SDB, there is substantial variation in the magnitude of that effect. The S-learner is optimized for settings where the overall treatment effect is zero (as we observe in this case) and indicates the largest reduction in SDB from the fully random design. The T-learner is better adapted to settings where there is a large treatment effect, but still provides results suggestive of a reduction in SDB from the use of conjoint analysis. The X-learner is instead aimed at observational settings where there control and treatment groups are imbalanced. While the X-learner should perform comparably to the T-learner in very large samples, it includes additional modeling to adjust for such imbalance but may be less efficient in smaller samples. Since our treatment was experimentally randomized, such imbalance is not present. Given the relatively small size of our sample (relative to those

growing each tree to .9, the number of trees used in each forest to 20,000, and set the minimum node size for observations in the averaging set when growing trees to 5. We also reduce the minimum node size in the splitting set to 1 for T and S learners to better match the behavior of the causal forest estimator, but leave this parameter at its default value for the X-learner which involves a more complex modeling procedure that does not as easily compare to the causal forest algorithm. Similarly, because causal forest algorithm imposes no absolute minimum on the node size (the minimum is instead implemented probablistically), we also alter the meta-learner arguments to impose no such strict minimum.

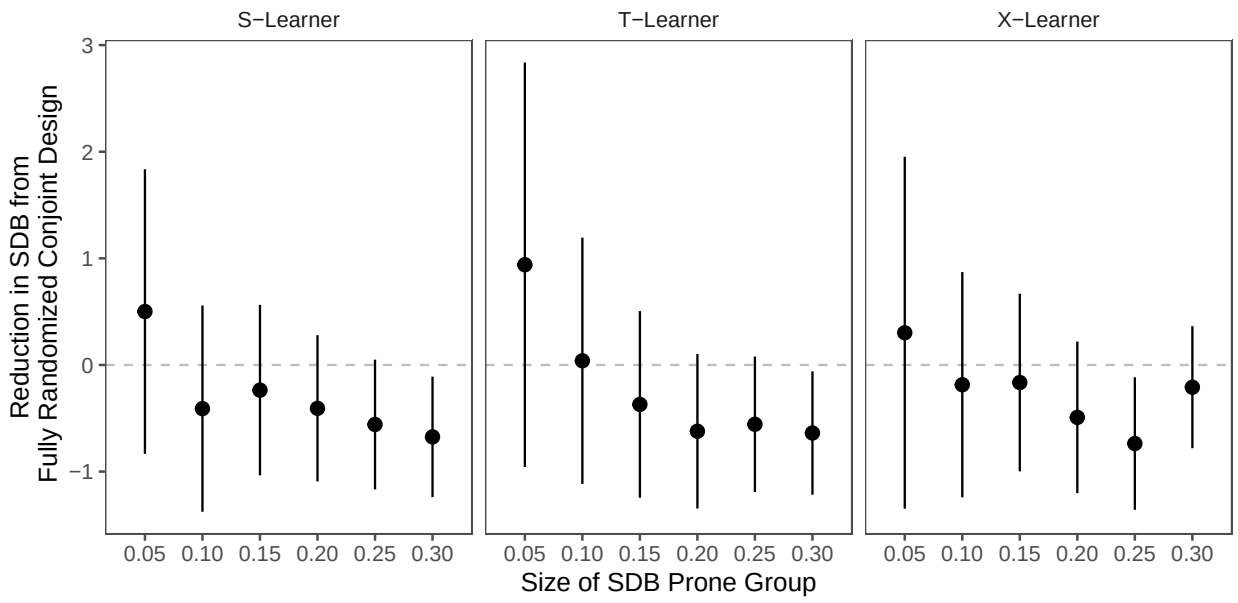

Figure D.3: Estimates For Reduction in SDB by Size of SDB Prone Group Using Meta-Algorithms

often used with flexible machine learning mechanisms), it is unsurprising that the algorithm performs relatively poorly.

D.9 Results using the Full Sample

In the pre-analysis plan, we registered the difference-in-differences estimation on a full sample as our primary analysis, in the hope that SDB would be strong enough for a large enough fraction of the sample to allow us to estimate a statistically significant reduction in SDB in such an analysis.

However, as Figure D.4 shows, the SDB-mitigating effect becomes virtually null when we use all respondents. Without the experimental stimulus expected to increase the consideration of social desirability, the AMCEs are -2.63 ($[-2.79, -2.46]$) under the partially randomized condition and -2.17 ($[-2.32, -2.03]$) under the fully randomized conditions. As expected, the design effect is negative: respondents are less likely to choose a candidate facing a sexual harassment scandal if the only attribute that varies is *Scandal*. However, not

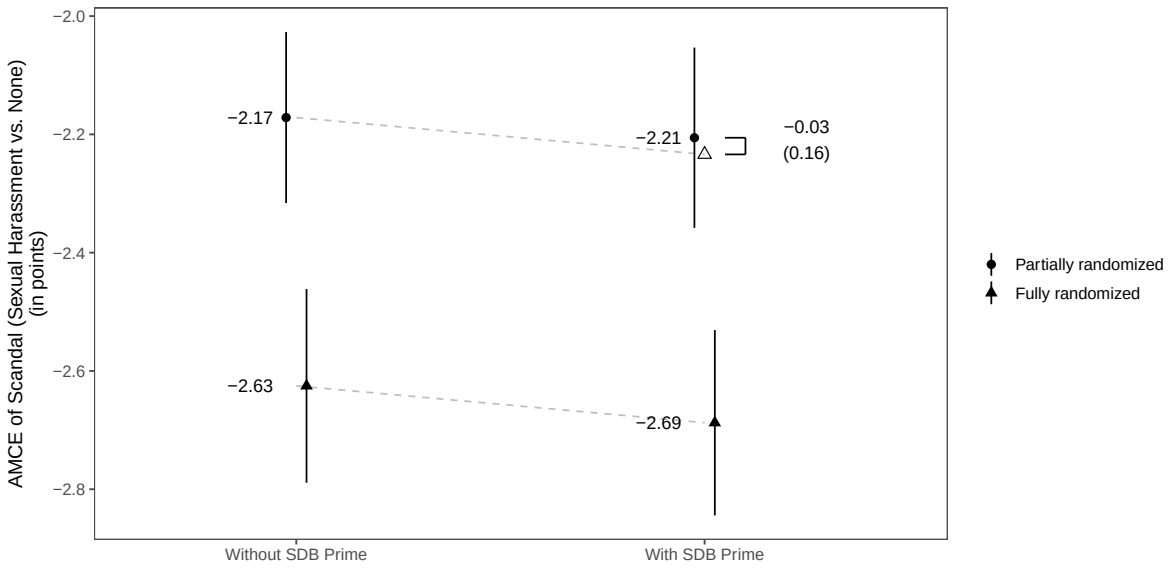

Figure D.4: Average Marginal Component Effects (AMCEs) for the Sensitive Attribute under the Partial and Fully Randomization Designs with and without the SDB-inducing Prime in Study 2. *See the caption for Figure 2 for the explanations of the graph elements. The outcome variable is a 7-point Likert Scale measure of preference for hypothetical candidates (least likely to vote for = 1; most likely to vote for = 7). This figure uses all respondents, including respondents identifies as not sensitive to SDB.*

in line with our original expectation, even with the prime, the AMCEs are similar: -2.69 ($[-2.84, -2.53]$) under the partially randomized condition and -2.21 ($[-2.36, -2.05]$) under the fully randomized condition. The difference in differences, our estimate of SDB-mitigating effect after subtracting the design effect, is -0.03 ($[-0.34, 0.28]$). Thus, for the whole sample, the prime fails to alter the AMCE of the scandal attribute neither in the fully random nor in the partial random conditions. We consider these results to indicate that the scandal attribute is not sensitive for most of the respondents in our sample. We therefore focus our analysis on the subset of the sample for which the prime appeared to have to successfully prime SDB.

With respect to heterogeneity across respondent subgroups, we pre-registered an intention to explore heterogeneity in the effect of the prime based on pre-treatment demographic

covariates without specifying the exact procedure. Because of this deviation from our pre-analysis plan, we consider our main analysis to be exploratory. That said, as discussed in Section 6.2, we use an algorithm that is specifically designed to "allow researchers to identify heterogeneity in treatment effects that was not specified in a preanalysis plan, without concern about invalidating inference due to searching over many possible partitions" (Athey and Imbens 2016, p.7353).

References for Appendix

Abrajano, Marisa A., Christopher S. Elmendorf and Kevin M. Quinn. 2018. "Labels vs. Pictures: Treatment-Mode Effects in Experiments About Discrimination." *Political Analysis* 26(1):20–33.

Adams, Troy L., Yuanxia Li and Hao Liu. 2020. "A Replication of Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research–Sometimes Preferable to Student Groups." *AIS Transactions on Replication Research* 6(1):15.

Athey, Susan and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113(27):7353–7360.

Athey, Susan, Julie Tibshirani and Stefan Wager. 2019. "Generalized Random Forests." *The Annals of Statistics* 47(2):1148–1178.

Kennedy, Ryan, Scott Clifford, Tyler Burleigh, Philip D. Waggoner, Ryan Jewell and Nicholas JG Winter. 2020. "The Shape of and Solutions to the MTurk Quality Crisis." *Political Science Research and Methods* 8(4):614–629.

Kinder, Donald R. 2013. Prejudice and Politics. In *Oxford Handbook of Political Psychology*, ed. Leonie Huddy, David O. Sears and Jack Levy. New York, NY: Oxford University Press.

Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel and Bin Yu. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning." *Proceedings of the National Academy of Sciences* 116(10):4156–4165.

Künzel, Sören R, Simon JS Walter and Jasjeet S Sekhon. 2019. "Causaltoolbox—Estimator Stability for Heterogeneous Treatment Effects." arXiv preprint, arXiv:1811.02833.

Mendelberg, Tali. 2001. *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality.* Princeton, NJ: Princeton University Press.

Montgomery, Jacob M and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62(3):729–744.

Nie, Xinkun and Stefan Wager. 2020. "Quasi-Oracle Estimation of Heterogeneous Treatment Effects." *Biometrika*, forthcoming.

Palan, Stefan and Christian Schitter. 2018. "Prolific.ac–A Subject Pool for Online Experiments." *Journal of Behavioral and Experimental Finance* 17:22–27.

Peyton, Kyle, Gregory A Huber and Alexander Coppock. 2020. "The Generalizability of Online Experiments Conducted During the COVID-19 Pandemic." SocArXiv Paper, `10.31235/osf.io/s45yg`.

Probst, Philipp and Anne-Laure Boulesteix. 2017. "To Tune or Not to Tune the Number of Trees in Random Forest." *The Journal of Machine Learning Research* 18(1):6673–6690.

Schwarz, Susanne and Alexander Coppock. 2020. "What Have We Learned About Gender From Candidate Choice Experiments? A Meta-analysis of 42 Factorial Survey Experiments." *The Journal of Politics*, forthcoming.

Teele, Dawn Langan, Joshua Kalla and Frances Rosenbluth. 2018. "The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics." *American Political Science Review* 112(3):525–541.

Tesler, Michael. 2016. *Post-Racial or Most-Racial? Race and Politics in the Obama Era*. Chicago, IL: University of Chicago Press.

Tibshirani, Julie, Susan Athey, Rina Friedberg, Vitor Hadad, David Hirshberg, Luke Miner, Erik Sverdrup, Stefan Wager and Marvin Wright. 2020. "Package `grf`: Generalized Random Forests." Version 1.2.0, available at the Comprehensive R Archive Network.

Valentino, Nicholas A., Fabian G. Neuner and L. Matthew Vandenbroek. 2018. "The Changing Norms of Racial Political Rhetoric and the End of Racial Priming." *The Journal of Politics* 80(3):757–771.

Wager, Stefan and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113(523):1228–1242.