Does Conjoint Analysis Mitigate Social Desirability Bias?

Yusaku Horiuchi* Zachary Markovich† Teppei Yamamoto‡

December 24, 2018

*Professor of Government and Mitsui Professor of Japanese Studies, Department of Government, Dartmouth College, 204 Silsby Hall, HB 6108, Hanover, NH 03755. Email: yusaku.horiuchi@dartmouth.edu, URL: https://sites.dartmouth.edu/horiuchi/

†Ph.D. Student, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: zmarko@mit.edu

‡Associate Professor, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: teppei@mit.edu, URL: http://web.mit.edu/teppei/www

Abstract

How can we elicit truthful responses in survey research? Scholars are often concerned about systematic mismeasurement in survey questions asking about sensitive topics due to social desirability bias (SDB). Recently, conjoint analysis has become a popular tool for measuring preferences when SDB is a concern, despite the lack of systematic evidence. In this paper, we employ a novel experimental design to investigate whether a standard fully randomized conjoint design can reduce bias in self-reported preferences about a socially sensitive behavior. We hypothesize that conjoint analysis mitigates SDB through two mechanisms, which we call imperception and rationalization. Our experiment isolates the SDB reduction through these mechanisms by comparing a standard conjoint design against a partially randomized design where only the socially sensitive attribute is randomly varied between the two profiles in each paired evaluation task. Our experiment also includes control conditions that are designed to remove confounding due to the increased attention to the varying attribute under the partial design. We implement the proposed experiment in an online survey about eco-friendly materials used in shoes. We find suggestive evidence that conjoint analysis does, indeed, mitigate SDB.

Keywords: response bias, survey experiment, survey methodology, conjoint analysis

1 Introduction

How can we elicit truthful responses in survey research? Scholars investigating socially undesirable behavior or politically sensitive topics are often concerned whether their survey measurement may suffer from social desirability bias (SDB) — systematic misreporting of socially sensitive behavior or attitudes toward "hot-button" issues, such as racial prejudice, drug use, and environmental concerns. Along with various survey techniques that have been developed to elicit honest answers, conjoint analysis has recently become increasingly popular among social scientists as a means to mitigate SDB (e.g., Hainmueller et al., 2014; Abrajano et al., 2015; Oliveros and Schuster, 2018; Flores and Schachter, 2018). However, there is little systematic evidence in the existing literature showing the effectiveness of conjoint analysis in reducing SDB in survey responses.

Conjoint analysis originates in marketing research (Green and Rao, 1971) and has been widely used for analyzing consumer behavior. Recently, Hainmueller et al. (2014) introduced the technique to the broader social science audience as a survey experimental method for causal inference based on the potential outcomes framework (Neyman, 1923; Rubin, 1974). This sparked numerous applications of conjoint analysis in the social sciences, particularly in political science, leading to a plethora of empirical studies on individuals' multidimensional preferences. To name a few examples, the method has been applied to study preferences about policy issues (e.g., Ballard-Rosa et al., 2017; Bechtel and Scheve, 2013; Horiuchi et al., 2018; Bansak et al., 2016; Flores and Schachter, 2018; Hainmueller and Hopkins, 2015; Schachter, 2016; Wright et al., 2016), bureaucracy (Oliveros and Schuster, 2018), politicians (e.g., Carnes and Lupu, 2016; Horiuchi et al., 2019; Teele et al., 2018), marketing (Green and Rao, 1971), the media (Helfer, 2016; Knudsen and Johannesson, 2018; Mummolo, 2016), censorship (Shao, 2018), terrorism (Huff and Kertzer, 2018), residential segregation (Mummolo and Nall, 2017), international trade (Bernauer and Nguyen, 2015; Spilker et al., 2016)), and NIMBYism (Hankinson, 2018).

In a standard conjoint experiment, respondents are presented with a table of two hypothetical profiles — for example, hypothetical candidates for faculty recruitment at a university (Carey et al., 2018; Polga-Hecimovich et al., 2019) — and asked to either choose the preferred profile or rate each. Each profile consists of a set of attributes that might affect respondents' preferences between the two profiles, such as candidates' race and gender. The attributes are randomly varied to form a series of pairwise comparisons. The resulting choice or rating responses are then aggregated to

estimate respondents' preferences about the attributes as well as their relative importance for the overall preferability of the profiles.

In this paper, we use a novel experimental design to investigate systematically whether fully randomized conjoint analysis can indeed mitigate SDB in survey-based measurement of preferences about sensitive attributes. Drawing on the theoretical literature, we hypothesize that conjoint analysis reduces SDB via two mechanisms (Section 2). First, respondents are less likely to perceive the sensitive nature of an attribute in a randomized conjoint experiment than when they are directly asked about the attribute because of the host of other attributes that are also randomized. Second, even if respondents recognize the potential of norm violation, they are more likely to express their honest preference because the other attributes enable them to rationalize their evaluations.

Our proposed design uses two pairs of parallel experiments to isolate the reduction in SDB through these mechanisms (Section 3). In the first pair of experiments, respondents are asked to express preferences about a sensitive attribute through one of the two types of conjoint evaluation tasks. In the fully randomized condition, respondents rate a series of paired profiles consisting of randomly assigned attributes as in a standard conjoint experiment. In the partially randomized condition, respondents are instead shown a series of paired profiles that are identical except for the sensitive attribute, thereby removing the SDB-mitigating mechanisms from the conjoint tasks while holding the marginal distributions of the attributes for each profile constant. This design is exactly replicated in the second pair of experiments, except that the sensitive attribute is replaced with a socially non-sensitive control attribute. The two pairs of experiments are then compared to calculate the change in the effect of the socially sensitive attribute between the fully random and partially random conditions, net of any additional change in the effects merely due to the increased attention to the varying attribute.

We implement our experimental design in an original two-wave survey experiment fielded online (Section 4). In the experiment, we ask respondents' likeliness to purchase a pair of athletic shoes with various attributes through 20 paired conjoint evaluation tasks. The treatment, socially sensitive attribute, is whether a pair of shoes is made of eco-friendly materials, whereas the control attribute is the use of gel-cushioning. Existing studies provide evidence that some people exhibit SDB in reporting their consciousness about environmental protection and eco-friendly behavior

(Auger and Devinney, 2007; Carrigan and Attalla, 2001; Carrington et al., 2010), whereas there is no reason to believe that gel-cushioning might induce such bias. Our statistical analysis shows suggestive evidence that the fully randomized conjoint design indeed reduces SDB among the pre-specified subsample of respondents who we identify as prone to SDB, while we find no such evidence for the rest of the sample (Section 5). Finally, we discuss possible limitations of our study as well as future research questions (Section 6).

2 Social Desirability Bias and Conjoint Analysis

A distinctive feature of conjoint experiments is that they measure respondents' preferences about multiple randomly generated attributes through a series of comparative evaluation tasks. Because of this design feature, we might expect conjoint experiments to mitigate SDB in preference measurement even when some of the attributes are "socially sensitive," compared to an alternative survey design which asks about respondents' preference for such a sensitive attribute directly. In other words, conjoint analysis might help mitigate SDB by allowing respondents to avoid directly stating their preferences about a socially sensitive attribute.

Theories regarding psychological mechanisms behind SDB often emphasize the costs that respondents face from giving non-socially desirable answers (Krumpal, 2013). These theories typically model survey responses as emerging from a cost-benefit analysis in which respondents balance the reward of positive emotions associated with honesty and helping researchers against embarrassment from giving certain answers and fears that the privacy of their answers may not be perfectly respected (Holtgraves, 2004). A strand of this literature relying on the rational choice theory suggests that respondents give inaccurate responses out of a rational concern that the privacy of their responses may not be fully protected and that they may face social sanctions from non-socially desirable answers (Stocké and Hunkler, 2007), while another strand of the literature emphasizes the cost due to embarrassment, which can be incurred even if respondents have perfect confidence in the anonymity of their answers (Rasinski et al., 1999).

The existing theories about the origin of SDB point at two possible mechanisms through which the conjoint survey design can mitigate SDB. First, both rational and subjective theories of SDB presuppose that respondents become aware of the possibility of violating social norms

and consciously avoid norm-violating responses. In a conjoint experiment, however, the sensitive attribute is included in the evaluation task along with a host of other non-sensitive attributes that are also randomly varied from task to task. Respondents are therefore unlikely to recognize the possibility of violating social norms by answering the conjoint questions in certain ways. We hypothesize that conjoint analysis mitigates SDB by this *imperception* mechanism.

Second, even if respondents become aware of the sensitive nature of one particular attribute, they are more likely to express their truthful preferences about the attribute because the other attributes enable them to rationalize their evaluations without explicitly violating social norms. The possibility of rationalization is likely to reduce the respondents' subjective cost of norm-violating responses whether their concerns derive from a rational calculation of the risk of privacy violation or fear of embarrassment. Thus, we hypothesize that conjoint analysis also mitigates SDB by this *rationalization* mechanism.

Existing research shows that many experimental designs which target these mechanisms have been successful at reducing SDB (Van Der Heijden et al., 2000; Lensvelt-Mulders et al., 2005; Holbrook and Krosnick, 2009; Cruyff et al., 2007; Coutts and Jann, 2011; Jann et al., 2011; Fisher, 1993). For example, researchers using the unmatched count technique ask respondents to report the total number of behaviors they have engaged in, some of which are socially sensitive and some of which are not, and can then recover the fraction that have engaged in the socially sensitive behavior by comparing the reported counts to answers to other questions about the non-sensitive behaviors (Holbrook and Krosnick, 2010; Wimbush and Dalton, 1997; LaBrie and Earleywine, 2000; Rayburn et al., 2003). Similarly, designs which ask respondent preferences regarding a series of profiles (e.g. factorial designs, indirect questioning, etc.) avoid asking respondents about the socially sensitive aspect of the profiles directly and so can help mitigate SDB (Wallander, 2009; Arnold and Feldman, 1981; Gonzalez-Ocantos et al., 2012).

3 Empirical Strategy

We design our study to estimate SDB reduction in conjoint experiments – or more specifically, the size of the response bias that is avoided in conjoint experiments thanks to the imperception and rationalization mechanisms. We do so by randomly assigning survey respondents into four

alternative conjoint survey experiments, or "design conditions," each of which is identically designed except in terms of how attributes are generated. In each design condition, respondents are asked to complete a series of paired conjoint evaluation tasks with multiple attributes characterizing profiles. Of these attributes, one is the *treatment attribute* which is likely to induce SDB in respondents' stated preferences. Another is the *control attribute* which is known to (or can safely be assumed to) induce no SDB in responses. The remaining attributes in the design are *filler attributes* which are not directly used in the analyses but included so as to activate the imperception and rationalization mechanisms in responses, as well as to maintain the realism of the conjoint tasks.

The four design conditions constitute a two-by-two factorial design, varying in two dimensions. Table 1 summarizes the four design conditions. First, the *partial randomization* designs randomly generate profiles in such a way that all but one attribute are identical within each pair of profiles. That is, respondents in the constrained designs are shown a series of profile pairs that are each identical in all but one attribute. In the *partial-treatment* (PT) condition, profiles vary only in terms of the treatment attribute, i.e., the socially sensitive attribute, whereas in the *partial-control* (PC) condition profiles differ in terms of the control, socially non-sensitive attribute.

Second, the *full randomization* designs randomly generate attributes much like a standard fully randomized conjoint experiment. That is, attributes for the two profiles in each pair are randomized independently from each other, resulting in a series of pairs that almost always differ in more than one attribute. However, the key difference from a typical, truly fully randomized conjoint experiment is that the randomization distribution is constrained so that the two profiles within each pair will always differ in either the treatment attribute (the *full-treatment* (FT) condition) or the control attribute (the *full-control* (FC) condition), so as to maintain the comparability of the resulting estimands with the partial randomization designs.

Formally, denote the joint probability mass function for the randomized attribute assignment for each paired conjoint task for design condition d by $f_d(t_1, c_1, a_1, t_2, c_2, a_2) \equiv \Pr(T_1 = t_1, C_1 = c, A_1 = a_1, T_2 = t_2, C_2 = c_2, A_2 = a_2 \mid D = d)$ for $d \in \{PT, PC, FT, FC\}$, where T_j, C_j and A_j represent the treatment attribute, control attribute, and filler attributes for profile $j \in \{1, 2\}$, respectively, and D represents the design condition for the respondent. Our experiment satisfies the following conditions:

Attribute That Always Varies	**Attribute Assignment Distribution**	
	Partial Randomization	Full Randomization
Treatment (Sensitive)	"PT" • Treatment attribute always different • Other attributes identical	"FT" • Treatment attribute always different • Other attributes fully randomized
Control (Non-sensitive)	"PC" • Control attribute always different • Other attributes identical	"FC" • Control attribute always different • Other attributes fully randomized

Table 1: Summary of the Design Conditions.

- $f_{PT}(t_1, c_1, a_1, t_2, c_2, a_2) = 0$ if $t_1 = t_2$, $c_1 \neq c_2$, or $a_1 \neq a_2$;
- $f_{PC}(t_1, c_1, a_1, t_2, c_2, a_2) = 0$ if $t_1 \neq t_2$, $c_1 = c_2$, or $a_1 \neq a_2$;
- $f_{FT}(t_1, c_1, a_1, t_2, c_2, a_2) = 0$ if $t_1 = t_2$;
- $f_{FC}(t_1, c_1, a_1, t_2, c_2, a_2) = 0$ if $c_1 = c_2$;
- $f_d(t_j, c_j, a_j) = f_{d'}(t_{j'}, c_{j'}, a_{j'})$ for all $d, d' \in \{PT, PC, FT, FC\}$ and $j, j' \in \{1, 2\}$; and
- $f_d(t_j, c_j, a_j) = f_d(t_j) f_d(c_j) f_d(a_j)$ for $d \in \{PT, PC, FT, FC\}$ and $j \in \{1, 2\}$, where $f_d(t_j)$, $f_d(c_j)$ and $f_d(a_j)$ denote the marginal probabilities for the treatment, control and filler attributes given $D = d$, respectively.

Our survey design ensures that individual profiles are identically distributed between the PT and FT conditions (and between the PC and FC conditions). This implies that any difference in the average marginal component effect (AMCE) of the treatment attribute between the PT and FT conditions must be attributed to difference in the covariance of the other attributes between the two profiles in each task. Specifically, in the PT condition, respondents always see a pair of profiles that are identical in all but the treatment attribute. We assume that this removes the imperception and rationalization mechanisms from the conjoint experiment. That is, because in every comparison task the socially sensitive attribute is the only difference between the two profiles, respondents are likely to perceive the task to be primarily purported to measure their preferences about that attribute (thereby removing imperception) and they are no longer able to justify their norm-violating evaluations by other attributes (thereby removing rationalization).

There are, however, two additional confounding factors that need to be considered before we can use the partial and fully random conditions to isolate SDB. First, in addition to the mechanisms that underlie SDB, the partial randomization is also likely to alter the AMCE of a given attribute through another mechanism unrelated to SDB, which we call a *design effect*. The design effect for the partially randomized conjoint design refers to the increase in the AMCE of the varying attribute due to the forced attention on that attribute. That is, in the partial randomization design, respondents are asked to evaluate the relative desirability of profiles solely based on the single attribute that varies within each pairwise comparison. What this implies is that even respondents who would put little weight on the attribute if other attributes were also varied between paired profiles are effectively forced to use the attribute as the only basis of their comparison, thereby amplifying its average effect across respondents. Due to the design effect, we expect the AMCEs to be larger in the partial randomization conditions than in the full randomization conditions for both the treatment attribute and the control attribute.

Second, respondents are not uniformly subject to SDB. The literature suggests that only certain types of survey respondents misreport their preferences due to social desirability concerns. Indeed, our empirical strategy hinges on the assumption that respondents would hesitate to report their socially undesirable preferences if they were asked about their preferences directly via a standard, non-conjoint survey question. In other words, we exclude respondents who are willing to openly admit they have a socially undesirable preference, and focus on "SDB-prone" respondents for our main analysis.

Our strategy for isolating SDB thus entails two steps. First, prior to the conjoint evaluation tasks, we directly asked respondents about their preferences on a socially sensitive object or behavior, which is to be used as the treatment attribute in the subsequent conjoint tasks. Respondents who are willing to express socially undesirable preferences in the direct question are then identified as "SDB-proof" respondents for this attribute and excluded from the main empirical analysis, since their conjoint responses cannot be affected by SDB even under the partial randomization condition. As discussed in Section 4, we asked these direct preference questions in a separate, pre-treatment wave that occured one week prior to the main survey, so that respondents' conjoint responses would not be affected by those questions.[1]

[1]In our empirical study, we also use a standard battery of items designed to measure proneness to SDB to

Second, using the remaining respondents who have been identified as potentially prone to SDB,[2] we compare the four design conditions and isolate out the portion of the change in the AMCE of the treatment attribute that is due to SDB. Specifically, we begin by estimating the AMCE among the SDB-prone respondents for the attribute that is always varied under each of the design conditions – the treatment attribute for the PT and FT conditions and the control attribute under the PC and FC conditions, to be more specific. We then calculate the increases in the AMCEs for these attributes when going from the fully randomized designs to the partially randomized designs. Finally, the discrepancy between these two increases becomes our estimate of the SDB reduction that can be attributed to conjoint experiments. Implicit in this strategy is the important assumption that the design effects are equal for the treatment and control attributes. That is, we assume that the treatment and control attributes are uniformly subject to the design effect, or the artificial increase in AMCEs due to the forced attention to those attributes under the partial designs. In Section 4, we discuss how we design our conjoint tasks to enhance the plausibility of this assumption; in Section 5, we empirically investigate possible threats to the validity of this assumption via robustness checks.

4 Survey Design

We employ our empirical strategy in an original survey experiment. The primary component of the experiment is a series of conjoint evaluation tasks on preferences about eco-friendly consumption behavior in the context of online shoe shopping. A key consideration in implementing our empirical strategy is to maximize the plausibility of our identification assumptions (in particular, the "equal design effect" assumption). An equally important element of our design choice is external validity: we want our fully randomized designs to be representative of the conjoint survey experiments typically implemented in political science and related fields of social sciences. The design of our

further subset the analysis sample. Details are discussed in Section 4.

[2]We note that not all of these respondents are prone to SDB because some of them may genuinely have a socially desirable preference about the treatment attribute. Our analysis sample may therefore still include respondents who do not belong to our population of interest. Our estimate of the reduction in SDB should thus be regarded as a lower bound for the true potential of a conjoint experiment for mitigating SDB.

two-wave survey experiment, described as follows, reflects these considerations.[3]

4.1 Wave 1

The first wave of our survey consists of descriptive measurement of respondents' demographic attributes and political attitudes, as well as a battery of items to measure proneness to SDB in terms of eco-friendly consumption behavior. We opt to measure these variables in a distal, separate pre-treatment wave despite the increased cost, so as to avoid priming respondents about the purpose of the study immediately before they answer the conjoint tasks. Making respondents aware of our primary interest in eco-friendly consumption preferences or SDB would draw their attention to our treatment attribute regardless of the design conditions, undermining the fundamental assumptions required for our empirical strategy.

To further ensure the validity of our empirical strategy, we present our survey to the respondents ostensibly as a survey about online shoe-shopping in general, not about buying eco-friendly shoes specifically. In addition, we mix our true questions of interest – opinions about environmental issues and a standard battery of items measuring proneness to SDB – with distracter questions about other attributes to be included in the conjoint experiment in Wave 2 (e.g. shoes brands, Amazon reviews) to disguise the main purpose of the study.

Our battery of SDB proneness questions are drawn from the Balanced Inventory of Desirable Responding Short Form (BIDR-16, Hart et al., 2015). Specifically, we will use eight questions that focus on a respondent's propensity towards impression management, but will exclude questions focused on measuring the propensity towards self-deception. We consider impression management, which is a respondent's tendency to lie in order to please others, as more relevant for our research than a respondent's tendency to give honest answers that do not reflect their real world behavior.[4]

[3] A detailed pre-analysis plan for our study was registered at the EGAP Design Registry prior to the launch of our Wave 1 survey, supplemented by an addendum filed between Wave 1 and Wave 2. The actual survey design and statistical analysis follows the pre-analysis plan exactly unless otherwise noted specifically. We note that a prototype of the proposed design had also been registered at the EGAP Registry and implemented separately, which informed the design of the current study.

[4] SDB is frequently divided into two varieties: impression management and self-deception (Paulhus and Reid, 1991). Impression management describes a respondent's tendency to present themselves in a more positive light towards others while self-deception represents a tendency to honestly respondent inaccurately. Self-deception arises when respondents maintain an unrealistically positive self-image and so will give inaccurate survey responses in order to help maintain that self-image. Conjoint analysis is likely to mitigate impression management most directly by asking questions in a way that minimizes the negative costs of providing socially undesirable answers.

We will also include one question from the longer BIDR-40 (Paulhus and Reid, 1991) that asks about a subject's propensity to litter, which we see as being directly relevant to a respondent's likelihood of stating a dishonest preference for an eco-friendly product.

Our Wave 1 survey thus consists of the following items: frequency of online shopping, experience of shopping shoes online, age, gender, marital status, race, partisanship, income, education, opinions about environmental issues, opinions about shoes, opinions about online shopping, and items from the BIDR battery. Of these covariates, we use the environment and BIDR variables to identify respondents who are potentially prone to SDB with respect to the treatment attribute, as discussed in Section 3. We fielded the survey on December 1 and 2, 2018, on 3,417 respondents recruited from Amazon's Mechanical Turk platform. The compensation for participation was \$0.60 per respondent. The median time to complete the survey was less than five minutes.

4.2 Wave 2

The second wave of our survey was conducted approximately one week after the conclusion of Wave 1 (December 8–15). We recontacted the respondents who had completed the first wave to be recruited to the second wave. The attrition rate turned out to be remarkably low (10%), yielding the final sample size of 3,075 for the two-wave study. We presume that the low attrition rate has been achieved at least partially due to the increased monetary incentive (\$1) for the second wave.

The assignment into the four design conditions was block randomized based on the covariates collected in Wave 1. 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. We use the following five covariates for the blocking: age ($\geq$ 36 years old vs. not), race (white vs. not), partisanship (Democrat, Republican, or other), opinion about environmental issues,[5] and proneness to SDB.[6] We then used each of the $2^4 \times 3 = 48$ unique cross-strata as a block, except the

[5]We code a respondent to be "anti-environmental" if he or she chooses the most anti-environmental option on any of the five items in the eco-friendliness battery.

[6]We first dichotomized each of the eight five-point-scaled SDB items such that the most and second most socially desirable options are coded as a "SDB-prone" response. We then coded a respondent as SDB-prone if he or she

Attribute	Levels
Gel Cushioning (**control**)	Has Gel Cushioning No Gel Cushioning
Eco-Friendly Materials (**treatment**)	100% Eco-Friendly Materials Used No Eco-Friendly Materials Used
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 2: List of Attributes for the Conjoint Experiment.

one block that contained only two respondents. We merged these respondents into a neighboring block (see Appendix A.1 for details), resulting in the total of 47 randomization blocks.

The Wave 2 survey consists of 20 paired conjoint evaluation tasks, which are identical for each of the four design conditions except the way conjoint profiles are generated (as discussed in Section 3). For each profile, we asked respondents how likely they are to purchase each of the two pairs of shoes using two separate 7-point Likert scales. We use this response measure instead of a more standard forced binary choice outcome because we expect the design effect to be larger if respondents were forced to make a binary choice, reducing the power of our test for detecting reduction in SDB. Table 2 shows the attributes for the profiles. Our treatment (i.e. socially sensitive) attribute is Eco-Friendly Materials, for which we expect a positive SDB for "100% Eco-Friendly Materials Used" against "No Eco-Friendly Materials Used" under the partial randomization design. We use Gel Cushioning as our control attribute, which takes on the same number of levels (i.e., two) as the treatment attribute, making the equal design effect assumption more likely to hold.

registered a SDB-prone response on four or more of the eight items.

5 Empirical Analysis

As we discussed in Section 3, our main analysis focuses on a subset of the respondents which we call the *SDB-prone respondents*. This group consists of respondents in our sample after excluding those for whom we expect no SDB about eco-friendly materials in shoes. We utilize two covariates measured in Wave 1 – opinions about environmental issues and BIDR – to identify those respondents. First, those who score low on the eco-friendliness variable are excluded from the main analysis sample. These questions directly ask about how much respondents care about the environment, so respondents who openly admit their lack of interest in protecting the environment should have no reason to express eco-friendly preferences in conjoint experiments. Second, the BIDR battery – our measure of proneness to SDB in general – are also used to exclude respondents who are unlikely to misreport their true, socially sensitive preferences because of SDB. More specifically, we define a respondent to be SDB-prone if the respondent is classified as pro-environment and SDB-prone based on the dichotomization we used for the block randomization (see Section 4). Note that using the same dichotomization rule for both block randomization and subset analysis ensures covariate balance between the design conditions within the analysis sample. After excluding the SDB-proof respondents and attrition, our analysis sample consists of 1,444 respondents (47% of the Wave 2 respondents).

Our main results are presented in the right panel of Figure 1. The plot presents our estimated AMCEs for the four design conditions (solid circles) along with their 95% confidence intervals based on standard errors robust to clustering at the respondent level (vertical bars).[7] Overall, the results are consistent with our expectation that the fully randomized conjoint design can reduce positive SDB in the measurement of preference for eco-friendly materials.

First, for the control attribute (gel cushioning), the AMCEs are substantially larger under the partially randomized design than the fully randomized design. That is, while the AMCE for gel

[7] We obtain these estimates via a variant of the least squared estimator proposed by Hainmueller et al. (2015) that incorporates the design 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 PT and FT conditions, and vice versa in the PC and FC 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 block dummies. The AMCE estimates can then be obtained as corresponding linear combinations of least squares coefficients on A1 and its interactions with the design dummies.

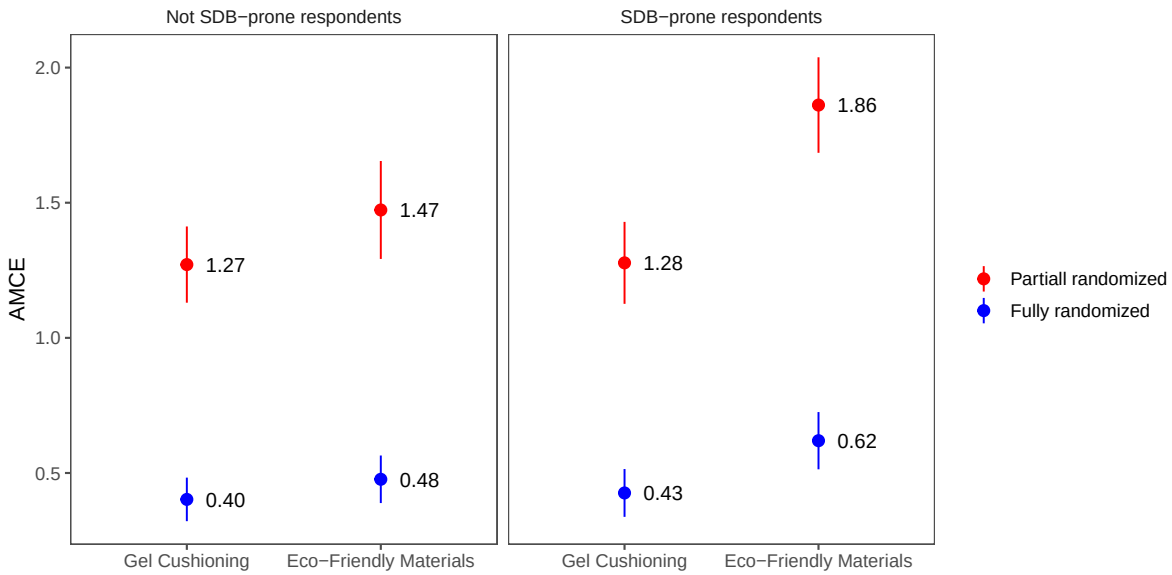

Figure 1: Average Marginal Component Effects (AMCEs) for the Treatment and Control Attributes under the Partially and Fully Randomized Designs. In each panel, the solid circles represent the estimated AMCEs for the four design conditions, and the vertical bars represent 95% confidence intervals based on standard errors robust to clustering at the respondent level.

cushioning (vs. no gel cushioning) is estimated to be 1.28 (with the 95% confidence interval of $[1.13, 1.43]$) on the 7-point scale outcome when only the gel cushioning attribute varies between the two shoe profiles in each pair, the AMCE for the same attribute drops to 0.43 ($[0.34, 0.51]$) when the other attributes are also varying. This shows that the design effect due to forced attention on a varying attribute in the partially randomized design is indeed substantial.

Second, although the same pattern holds for the treatment attribute (eco-friendly materials), the gap between the partial and full randomization designs appears to be even larger than for the control attribute. That is, the AMCE for shoes made of 100% eco-friendly materials (vs. no eco-friendly material) is estimated to be 1.86 (with the 95% confidence interval of $[1.68, 2.04]$) under the partially randomized design but it is only 0.62 ($[0.51, 0.73]$) under the fully randomized design.

The estimated difference between the two differences ($= 0.39$) is significantly different from zero at conventional levels of statistical significance, with the 95% confidence interval of $[0.12, 0.66]$. However, testing the null of no difference between the two ratios (i.e. the ratio of the partial-randomization AMCE to the full-randomization AMCE for each attribute) returns an insignificant

result.[8]

Remarkably, the pattern disappears when we conduct the same analysis on the respondents who we excluded from the main analysis, i.e., the "SDB-proof" subsample based on our direct measure of opinions about environmental issues as well as general proneness to SDB. That is, as presented in the left panel of Figure 1, the AMCE for the control attribute is 1.27 ($[1.13, 1.41]$) and 0.40 ($[0.32, 0.48]$) under the partial and full randomization conditions, respectively, which amounts to the difference of 0.87 between the two design conditions. The corresponding estimates for the treatment 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 is now statistically insignificant with the 95% confidence interval of $[-0.13, 0.39]$. The difference-in-ratios result also indicates there is no difference between the treatment and control attributes.

Our analysis so far indicates suggestive evidence that conjoint analysis does mitigate SDB in the measurement of preference about sensitive attributes. One possible concern for the validity of our study, however, is differential response bias across the design conditions other than SDB. In particular, survey fatigue might affect respondents in different design conditions differentially in such a way that confounds our estimate of SDB. For example, one might hypothesize that the partially randomized designs would induce respondents to satisfice more than the fully randomized designs because the tasks might feel more repetitive in the partially randomized conditions. If such differential satisficing between the partially and fully randomized conditions further interacts with the difference between the treatment and control attributes, our SDB estimate would be confounded by the difference in satisficing.

To test if differential satisficing is indeed a likely threat to our inference, we investigate whether the difference-in-differences in the AMCEs across the four design conditions might grow over the course of the 20 tasks each respondent completes. That is, a significant interaction between the difference-in-differences and the task count would indicate that there is evidence of differential fatigue across the conditions over the course of our survey. Figure 2 graphically presents the result of this test, implemented in two different ways. Neither of these suggests evidence of differential

[8]In our pre-analysis plan, we specified the difference-in-ratios test as the primary test procedure, without noting the possibility of conducting a difference-in-differences test as an alternative. The p-values reported here should therefore not be interpreted as confirmatory evidence against the null hypothesis. With hindsight, however, we consider the two procedures to be equally justifiable and we could well have specified the difference-in-differences as the main analysis.

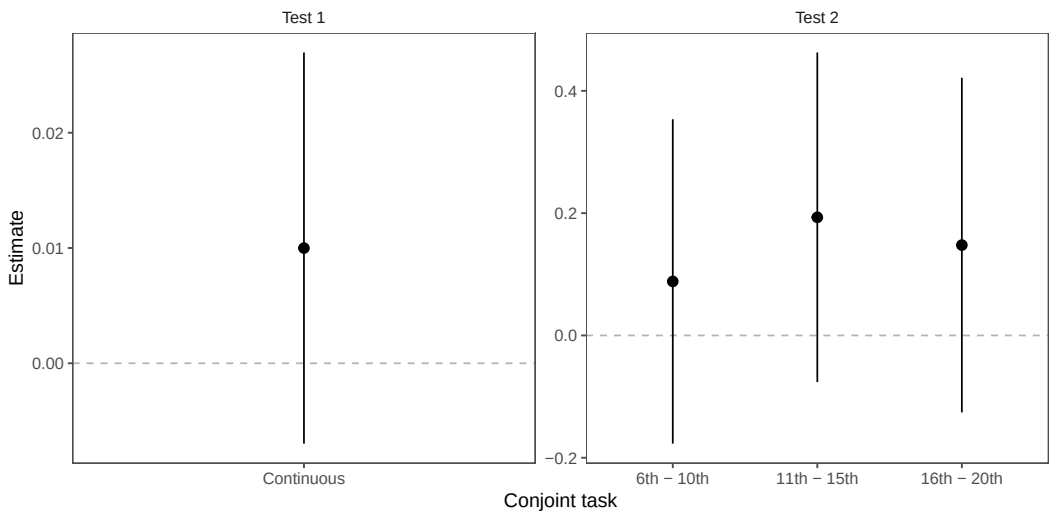

Figure 2: Robustness for Fatigue Effects.

satisficing. First, when the task count is treated as a continuous variable, its linear interaction with the difference-in-differences term (0.01) is statistically indistinguishable from zero, with a p-value of 0.25. Second, to account for a possible nonlinearity, we recode the task count into a four-level categorical variable (the 1st to 5th tasks, 6th to 10th tasks, 11th to 15th tasks, and 16th to 20th tasks) and estimate the interaction between each of the category dummies and the difference-in-differences term. None of the three interaction effects (0.09, 0.19, and 0.15) are statistically distinguishable from zero, with p-values of 0.51, 0.16, and 0.29, respectively. Thus, we conclude that differential fatigue is a highly unlikely threat to the validity of our primary analysis.

6 Conclusion

Conjoint analysis has become a popular tool for analyzing respondent preferences when SDB is a concern, but there is little empirical evidence that it is appropriate for this task. Because conjoint analysis avoids directly asking respondents about their preferences regarding a socially sensitive characteristic, there is a strong theoretical basis for believing that it will at least partially reduce SDB. Our empirical results are consistent with this belief. We observe that the change in effect sizes observed moving from a fully randomized conjoint design to a constrained conjoint design is larger for a socially sensitive attribute than a non-sensitive attribute. We theorize that the

difference in AMCE's between the two designs will be a combination of both the reduction in SDB and a design effect which emerges from focusing the respondent's attention on a single attribute. Consequently, the positive difference in differences suggests that conjoint analysis results in a reduction in SDB.

We implemented this design in the context of shopping for athletic shoes online. Specifically, the socially sensitive attribute was the use of eco-friendly materials and the non-sensitive attribute was the presence of gel cushioning. Political scientists are often interested in opinions surrounding the environment and worry that SDB could tarnish the measurement of these opinions, and conjoint analysis has been used in exactly this context (Bechtel and Scheve, 2013). Moreover, there is a known disconnect between how likely consumers say that they are to purchase eco-friendly materials in surveys and the aggregate sales data, suggesting that SDB is a significant concern in this domain (Auger and Devinney, 2007; Carrigan and Attalla, 2001; Carrington et al., 2010). We therefore consider this to be a useful area in which to test the efficacy of conjoint analysis.

Although a useful application domain, it will be important for future research to validate these findings in other areas. For example, the mechanisms that lead some respondents to overstate their propensity to purchase an eco-friendly product may be very different from the ones that lead them to underreport drug use. Still it should be possible to adapt our design to these settings. For example, conjoint analysis could be used to measure the preferences of college students about what kinds of parties they might attend. The presence or absence of alcohol or marijuana could be the socially sensitive attribute and free food being offered at the event could be the non-sensitive attribute.

Additionally, it will be important to examine how these effects vary based off of the nature of the conjoint task. We have considered the specific setting in which both the sensitive and non-sensitive attributes take on only two values and respondents rate both profiles on a scale of 1-7. Future research will be needed to determine whether the reduction in SDB also occurs for a forced choice experiment, in which respondents must choose which of the two profiles they prefer or when the attributes have more than two levels.

Still the results in this paper are encouraging. They provide empirical support for the already common practice of using conjoint analysis to elicit honest preferences regarding socially sensitive attributes. Moreover, the reduction in SDB we observe in this case is substantively large.

References

Abrajano, M. A., Elmendorf, C. S., and Quinn, K. M. (2015). Using experiments to estimate racially polarized voting. UC Davis Legal Studies Research Paper Series, No. 419.

Arnold, H. J. and Feldman, D. C. (1981). Social desirability response bias in self-report choice situations. Academy of Management Journal, 24(2):377–385.

Auger, P. and Devinney, T. M. (2007). Do what consumers say matter? the misalignment of preferences with unconstrained ethical intentions. Journal of Business Ethics, 76(4):361–383.

Ballard-Rosa, C., Martin, L., and Scheve, K. (2017). The structure of american income tax policy preferences. The Journal of Politics, 79(1):1–16.

Bansak, K., Hainmueller, J., and Hangartner, D. (2016). How economic, humanitarian, and religious concerns shape european attitudes toward asylum seekers. Science, page aag2147.

Bechtel, M. M. and Scheve, K. F. (2013). Mass support for global climate agreements depends on institutional design. Proceedings of the National Academy of Sciences, 110(34):13763–13768.

Bernauer, T. and Nguyen, Q. (2015). Free trade and/or environmental protection? Global Environmental Politics, 15(4):105–129.

Carey, J. M., Carman, K. R., Clayton, K. P., Horiuchi, Y., Htun, M., and Ortiz, B. (2018). Who wants to hire a more diverse faculty? a conjoint analysis of faculty and student preferences for gender and racial/ethnic diversity. Politics, Groups, and Identities, pages 1–19.

Carnes, N. and Lupu, N. (2016). Do voters dislike working-class candidates? voter biases and the descriptive underrepresentation of the working class. American Political Science Review, 110(4):832–844.

Carrigan, M. and Attalla, A. (2001). The myth of the ethical consumer–do ethics matter in purchase behaviour? Journal of consumer marketing, 18(7):560–578.

Carrington, M. J., Neville, B. A., and Whitwell, G. J. (2010). Why ethical consumers don't walk their talk: Towards a framework for understanding the gap between the ethical purchase

intentions and actual buying behaviour of ethically minded consumers. Journal of business ethics, 97(1):139–158.

Coutts, E. and Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (rrt) and the unmatched count technique (uct). Sociological Methods & Research, 40(1):169–193.

Cruyff, M. J., van den Hout, A., van der Heijden, P. G., and Böckenholt, U. (2007). Log-linear randomized-response models taking self-protective response behavior into account. Sociological methods & research, 36(2):266–282.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. Journal of consumer research, 20(2):303–315.

Flores, R. D. and Schachter, A. (2018). Who are the "illegals"? the social construction of illegality in the united states. American Sociological Review, 83(5):839–868.

Gonzalez-Ocantos, E., De Jonge, C. K., Meléndez, C., Osorio, J., and Nickerson, D. W. (2012). Vote buying and social desirability bias: Experimental evidence from nicaragua. American Journal of Political Science, 56(1):202–217.

Green, P. E. and Rao, V. R. (1971). Conjoint measurement for quantifying judgmental data. Journal of Marketing research, pages 355–363.

Hainmueller, J. and Hopkins, D. J. (2015). The hidden american immigration consensus: A conjoint analysis of attitudes toward immigrants. American Journal of Political Science, 59(3):529–548.

Hainmueller, J., Hopkins, D. J., and Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. Political Analysis, 22(1):1–30.

Hainmueller, J., Hopkins, D. J., and Yamamoto, T. (2015). Learning more from conjoint experiments through a doubly randomized design. Paper presented at the Annual Meeting of APSA.

Hankinson, M. (2018). When do renters behave like homeowners? high rent, price anxiety, and nimbyism. American Political Science Review, pages 1–21.

Hart, C. M., Ritchie, T. D., Hepper, E. G., and Gebauer, J. E. (2015). The balanced inventory of desirable responding short form (bidr-16). SAGE Open, 5(4):1–9.

Helfer, L. (2016). Media effects on politicians: An individual-level political agenda-setting experiment. The International Journal of Press/Politics, 21(2):233–252.

Holbrook, A. L. and Krosnick, J. A. (2009). Social desirability bias in voter turnout reports: Tests using the item count technique. Public Opinion Quarterly, 74(1):37–67.

Holbrook, A. L. and Krosnick, J. A. (2010). Measuring voter turnout by using the randomized response technique: Evidence calling into question the method's validity. Public Opinion Quarterly, 74(2):328–343.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. Personality and Social Psychology Bulletin, 30(2):161–172.

Horiuchi, Y., Smith, D. M., and Yamamoto, T. (2018). Measuring voters' multidimensional policy preferences with conjoint analysis: Application to japan's 2014 election. Political Analysis, 26(2):190–209.

Horiuchi, Y., Smith, D. M., and Yamamoto, T. (2019). Identifying voter preferences for politicians' personal attributes: A conjoint experiment in japan. PoliticalScience Research and Method, page forthcoming.

Huff, C. and Kertzer, J. D. (2018). How the public defines terrorism. American Journal of Political Science, 62(1):55–71.

Jann, B., Jerke, J., and Krumpal, I. (2011). Asking sensitive questions using the crosswise model: an experimental survey measuring plagiarism. Public Opinion Quarterly, 76(1):32–49.

Knudsen, E. and Johannesson, M. P. (2018). Beyond the limits of survey experiments: How conjoint designs advance causal inference in political communication research. Political Communication, pages 1–13.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. Quality & Quantity, 47(4):2025–2047.

LaBrie, J. W. and Earleywine, M. (2000). Sexual risk behaviors and alcohol: Higher base rates revealed using the unmatched-count technique. Journal of Sex Research, 37(4):321–326.

Lensvelt-Mulders, G. J., Hox, J. J., Van der Heijden, P. G., and Maas, C. J. (2005). Meta-analysis of randomized response research: Thirty-five years of validation. Sociological Methods & Research, 33(3):319–348.

Mummolo, J. (2016). News from the other side: How topic relevance limits the prevalence of partisan selective exposure. The Journal of Politics, 78(3):763–773.

Mummolo, J. and Nall, C. (2017). Why partisans do not sort: The constraints on political segregation. The Journal of Politics, 79(1):45–59.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. (translated in 1990). Statistical Science, 5:465–480.

Oliveros, V. and Schuster, C. (2018). Merit, tenure, and bureaucratic behavior: Evidence from a conjoint experiment in the dominican republic. Comparative Political Studies, 51(6):759–792.

Paulhus, D. L. and Reid, D. B. (1991). Enhancement and denial in socially desirable responding. Journal of Personality and Social Psychology, 60(2):307.

Polga-Hecimovich, J., Carey, J. M., and Horiuchi, Y. (2019). Student attitudes toward campus diversity at the united states naval academy: Evidence from conjoint survey experiments. Armed Forces & Society, page forthcoming.

Rasinski, K. A., Willis, G. B., Baldwin, A. K., Yeh, W., and Lee, L. (1999). Methods of data collection, perceptions of risks and losses, and motivation to give truthful answers to sensitive survey questions. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 13(5):465–484.

Rayburn, N. R., Earleywine, M., and Davison, G. C. (2003). Base rates of hate crime victimization among college students. Journal of Interpersonal Violence, 18(10):1209–1221.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized-studies. Journal of Educational Psychology, 6:688–701.

Schachter, A. (2016). From "different" to "similar" an experimental approach to understanding assimilation. American Sociological Review, 81(5):981–1013.

Shao, L. (2018). The dilemma of criticism: Disentangling the determinants of media censorship in china. Journal of East Asian Studies, 18(3):279–297.

Spilker, G., Bernauer, T., and Umaña, V. (2016). Selecting partner countries for preferential trade agreements: Experimental evidence from costa rica, nicaragua, and vietnam. International Studies Quarterly, 60(4):706–718.

Stocké, V. and Hunkler, C. (2007). Measures of desirability beliefs and their validity as indicators for socially desirable responding. Field Methods, 19(3):313–336.

Teele, D. L., Kalla, J., and Rosenbluth, F. (2018). The ties that double bind: Social roles and women's underrepresentation in politics. American Political Science Review, pages 1–17.

Van Der Heijden, P. G., Van Gils, G., Bouts, J., and Hox, J. J. (2000). A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning: Eliciting sensitive information in the context of welfare and unemployment benefit. Sociological Methods & Research, 28(4):505–537.

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. Social Science Research, 38(3):505–520.

Wimbush, J. C. and Dalton, D. R. (1997). Base rate for employee theft: Convergence of multiple methods. Journal of Applied Psychology, 82(5):756.

Wright, M., Levy, M., and Citrin, J. (2016). Public attitudes toward immigration policy across the legal/illegal divide: The role of categorical and attribute-based decision-making. Political Behavior, 38(1):229–253.

Appendix

A.1 Additional Details of the Survey Design

We implemented a block randomization strategy to eliminate potential confounding from observed covariates. In Wave 1, we collected a large amount of demographic and political covariates about respondents. We constructed blocks based off of this information. Specifically, we blocked on age, race, partisanship, environmental attitudes, and SDB proneness. The subgroup that we were primarily interested in was both SDB prone and not anti-environment, so blocking on these variables ensured that treatment assignment was balanced. We chose the other three blocking covariates: age, race, and partisanship because substantively we believe that each of these is important for determining preference about athletic shoes.

When creating blocks, we coarsened age so that it represented whether or not a respondent was over 35 years old and race 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 categorized respondents as SDB prone if they registered an SDB prone response on four or more of the eight SDB questions that we posed (see Sections 4 and 5 for additional information).

Table A.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.

Table A.1: Block Randomization

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

Note: Block 11 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).