Social desirability bias in candidate conjoint experiments: What is the optimal design when studying sensitive topics?
Jacobsen, Malte Dahl

Publication date: 2018

Citation for published version (APA):
Jacobsen, M. D. (2018). Social desirability bias in candidate conjoint experiments: What is the optimal design when studying sensitive topics?
Social desirability bias in candidate conjoint experiments: What is the optimal design when studying sensitive topics?

Malte Dahl
September 2018

Abstract

An often-mentioned advantage over traditional survey experiments is that conjoint experiments have the potential to mitigate social desirability bias. Because respondents are presented with numerous features, the research objective is concealed and respondents can find multiple justifications for any given choice. However, if and to what extent this is true may depend on design choices – a concern that has received little empirical attention. In this paper, I randomly assign respondents to six conjoint designs in order to induce respondents’ awareness to sensitive features and their possibilities of acting strategically. The results show that while design variations has a substantial effect on respondents’ awareness to sensitive features, it has no detectable effect on their priorities. The evidence suggests that the fear of social desirability bias should not force researchers to use conjoint designs that are otherwise suboptimal.

Working paper presented at seminar on Political Behavior, UCPH September 21
1 Introduction

Candidate conjoint experiments have become a standard part of the political science toolkit. For example, conjoint designs have been used to study gender-bias towards political candidates (Teel, Kalla, and Rosenbluth 2018), how explicit partisan affiliation affect voters’ preferences (Kirkland and Coppock 2017), and attitudes toward immigrants (Hainmueller and Hopkins 2015). Conjoint experiments enable researchers to examine multidimensional preferences over candidates and allow for comparing the relative importance of these features. A considerable advantage that is also often emphasized, is that conjoint designs have the potential to mitigate social desirability bias (SDB) (Hainmueller, Hopkins, and Yamamoto 2014; Horiuchi, Smith, and Yamamoto 2017; Teel, Kalla, and Rosenbluth 2018). This is an essential quality considering that candidate conjoint experiments are often used to gauge respondents’ reactions to sensitive dimensions such as candidates’ race or sex. In fact, the usefulness of these studies depends on our ability to obtain valid self-reported answers. However, despite the prominence of conjoint analysis, little empirical work has paid attention to assessing to what extent and under what conditions SDB is in fact a concern in conjoint experiments.

Conjoint experiments come in a number of design varieties. First, there is a fundamental difference in terms of whether the candidate profiles are shown in pairs or not.\(^1\) Secondly, outcomes can be based on either a discrete choice, a rating of the candidates, or a combination of the two. Thirdly, the number of features (features can for example be Age/Sex/Party affiliation/Race) varies with studies displaying a few up to a dozen different features. Moreover, different randomization schemes can be applied, with some studies randomizing the number of features that are presented, randomizing all or only some features, and/or randomizing feature levels (e.g. Black/White/Hispanic) with different probability weights (See Hainmueller, Hopkins, and Yamamoto (2014) for a discussion). These design differences can seem subtle, but they might greatly affect (1) respondents’ inferences about the primary research objective, (2) their possibilities for acting strategically when answering questions, and (3) the degree to which respondents can justify inappropriate answers. All of which might cause respondents’ answers to be biased.

Survey respondents are often aware that they are being evaluated by an unseen but omnipresent scholar (Berinsky 2004), and it therefore matters how they perceive research intentions. It is a widely held belief that respondents’ inferences about the research objective affect how likely they are to reveal answers that are socially unaccepted. In other words, a systematic bias can be introduced if the purpose of the experiment is too obvious (McDermott 2002). For this reason, scholars often avoid revealing their intentions, or include strategies to deceive survey participants about the research agenda (McDermott 2002; Berinsky, Huber, and Lenz 2012).

\(^1\)The most common design is the paired conjoint experiment in which respondents are presented with two profiles at the same time, whereas in the single-profile conjoint, only one profile is displayed and evaluated at a time (see for example Sen (2017)).
The degree to which SDB is a concern in conjoint experiments may be conditioned on the design differences mentioned above. The paired-conjoint design arguably provides a more noticeable contrast between feature levels, for example by repeatedly asking respondents to choose between a black and a white candidate, than is the case in the single-profile conjoint where control or treatment is not displayed at the same time. Additionally, the within-subject structure of the paired design makes it easier for respondents to act strategically since they are presented with both treatment and control at the same time (Mummolo and Peterson 2018). Another important design component is the probability weight of sensitive feature levels. A high frequency of black candidates might raise respondents’ awareness about race as a primary research objective. Moreover, by asking respondents repeatedly to make a choice on a sensitive topic arguably reduces their possibilities of justifying socially inappropriate answers.

In this study, I randomly assign respondents to conjoint designs intended to either downplay or increase respondents’ attention to sensitive dimensions and their possibilities of acting strategically. Specifically, I examine how the paired structure and different probability weights of sensitive feature levels affect respondents’ anticipation of the research objective and consequently their priorities. This work is related to recent studies that examine demand effects in survey experiments by inducing different degrees of information about the purpose of the study (Mummolo and Peterson 2018; De Quidt, Haushofer, and Roth 2017). However, instead of raising awareness about the research objective by providing explicit information about it, the present project seek to manipulate awareness to sensitive dimensions through design. To do so, I run a total of six candidate conjoint experiments on two sensitive topics. The studies rely on convenience-samples and is conducted at Amazon’s Mechanical Turk (N = 7,059).

Overall, the conjoint experiments give remarkably similar results across designs. While there is a significant effect on the share of respondents that anticipate the sensitive feature as the main research intent, their evaluations of these features remain stable across conditions. More specifically, respondents assigned to a paired conjoint design with frequent contrast on a sensitive feature are on average much more likely to believe the sensitive dimension is the main focus of the study vis-à-vis conditions that provide less contrast. This should amplify SDB. However, and most importantly, it does not translate into any immediate effect on respondents’ priorities compared to the designs that succeeded in masking the sensitive feature. In none of the studies do the samples change the substantial interpretation of the sensitive dimension and in no instance did the experiments produce significantly distinguishable effects.

2 Social desirability bias in survey experiments

A common understanding of SDB is the respondent’s lack of comfort to reveal his or her true attitudes (Tourangeau and Yan 2007; Kumar 2012; Holtgraves 2004). Respondents moderate their behavior in order to make themselves look
more favourably and avoid the embarrassment, unease and distress that revealing socially undesirable answers may bring (Kaminska and Foulsham 2013). Research on SDB indicates that respondents tend to underreport favoritism for a preferred group of people as compared to a nonpreferred group of people (Janus 2010; Kuklinski et al. 1997). SDB bias thus masks the effect of in-group favoritism and lead to a misrepresentation of preferences. Evidence on SDB bias in survey studies in general suggests that it is a valid concern, which is proved by wording questions in less threatening ways (Kuklinski et al. 1997), by changing the interview setting (Krysan and Couper 2003), or by comparing list experiments with direct questions (Janus 2010; Gilens, Sniderman, and Kuklinski 1998). Finally, some studies have examined the relationship between racial attitudes and measures on self-monitoring, indicating that high self-monitors are more likely to offer socially acceptable answers (Berinsky and Lavine 2012).

A particular concern is that survey experiments frequently rely on online subject pools, like Amazon’s Mechanical Turk, where experienced experimental participants have incentives to be especially attentive to researcher expectations (Krupnikov and Levine 2014). For this reason, Berinsky, Huber, and Lenz (2012) recommend that researchers avoid revealing their intentions in online survey experiments.\textsuperscript{2}

\textbf{2.1 Conjoint experiments as a mean to overcome SDB}

While SDB is a potential validity issue in all survey experiments, it is often claimed that conjoint experiments can mitigate some of these concerns (Hainmueller, Hopkins, and Yamamoto 2014; Liu 2018; Teele, Kalla, and Rosenbluth 2018). This notion builds on two arguments. First, because research participants are typically presented with a large number of features, the design allows respondents to justify any particular choice or rating (Hainmueller, Hopkins, and Yamamoto 2014). Secondly, because of the large number of varying features, it is unclear to respondents what the main research objective of the study is (Hainmueller, Hopkins, and Yamamoto 2014; Ono and Yamada 2016). For example, in their conjoint experiment on the importance of gender stereotypes Teele, Kalla, and Rosenbluth (2018) state that because candidate sex is embedded as one of multiple features ’(...) our own interest in gender would not have been obvious in the experiment. This likely lessens the degree to which our results are skewed by social desirability bias’.

Hiding research intentions to mitigate SDB is not a new idea. Previous research on sensitive topics have implemented cover stories in order to misdirect participants about the goal of the experiment (McDermott 2002; Dickson 2011). For example, by asking questions unrelated to the primary intention of the study (Kam 2007) or by providing respondents with an alternative or vaguely stated purpose of the experiment (Bullock 2011; Arceneaux 2008).

\textsuperscript{2}Researcher demand effect is distinct from SDB and happens when respondents infer the response researchers expect and behave in line with these expectations (Mummolo and Peterson 2018). In principle, demand effects could work in the opposite direction of SDB.
While the arguments against response bias in conjoint experiments appear sound, there is little empirical evidence to support them. Previous candidate conjoint studies implement various tests in order to examine potential SDB. For example, Bansak, Hainmueller, and Hangartner (2016) find that results are constant across respondents with different levels of empathy, building on the idea that empathy and social desirability scales correlate. Hainmueller and Hopkins (2015) reestimate their results based on measures of self-monitoring that are known to be closely connected to social desirability. Finally, Hainmueller, Hangartner, and Yamamoto (2015) use a natural experiment as a behavioral benchmark and compare survey techniques with real-world behavior.

While these studies provide some reason not to be concerned about SDB, results from several candidate conjoint experiments on sensitive dimensions question the validity of answers to sensitive features. For example, a number of studies find positive effects of candidates being black (Carnes and Lupu 2016; Kirkland and Coppock 2017). Moreover, despite considerable study-to-study variation, a majority of conjoint experiments that include gender show a net preference for women (Schwarz, Hunt, and Coppock 2018). While this of course can reflect respondents true preferences, we might worry that these experiments overestimate support for female and black candidates. This notion is supported by recent evidence suggesting that experimental findings on voter preferences for women or black candidates may overestimate support, even in anonymous settings (Krupnikov, Piston, and Bauer 2016).

3 Research design

I conduct two studies each comprising three conjoint experiments specifically designed to assess the relation between design and SDB. Each study is substantively inspired by previous candidate conjoint experiments. The studies were fielded on a total of 7,059 respondents recruited from Amazon’s Mechanical Turk, which hosts an experienced pool of survey respondents (Berinsky, Huber, and Lenz 2012).

3.1 Manipulating attention to sensitive features through design

Both studies include a topic that is known to be influenced by social desirability pressures. Study 1 tries to gauge the effect of candidates race, a topic to which it can be difficult to obtain honest self-reports since racial preferences is taboo

\footnote{Carnes and Lupu (2016) conduct a conjoint experiment in which they manipulate candidates’ race using two levels (white or black) in a study of support for political candidates, and find a positive (although only borderline significant) effect of being black. In addition, Kirkland and Coppock (2017) finds that Hispanic, Black and Asian candidates respectively are preferred over White candidates.}

\footnote{The present study differ slightly from the original studies in terms of the number and type of features included. Since the purpose of the present study is to determine whether treatment effects vary across design, this is not an issue.}
Study 2 seeks to unravel support for immigrants seeking admission to the US. Religious affiliation, more specifically being Muslim, serve as a sensitive feature. Restrictionist immigration policies is a topic that previous research has found to be subject to response bias (Janus 2010) and with recent discussions in the U.S. about a ”Muslim ban”, this is a hotbutton topic that is likely to invoke social desirability pressures.

I seek to manipulate respondents awareness to the sensitive feature in two ways. First, by varying the probability weight of the sensitive feature across conditions. Thus, the baseline design is a paired-conjoint in which respondents are presented with five candidate pairs with a frequent contrast on the sensitive level. Arguably it also makes it harder for respondents to defend a ”bad” answer repeated across five pairs. In the second condition, the sensitive level appear less frequently ”masking” it from respondents.

Secondly, using a single-profile in stead of a paired conjoint, the sensitive feature will arguably appear less noticeable. While respondents in the paired conjoint designs observe both treatment and control at the same time, in the single-profile conjoint they receive either control or treatment. Arguably, this also makes it harder for respondents to act strategically (Mummolo and Peterson 2018). The single-profile conjoint design is dissimilar to the paired designs in the sense, that respondents can not make a binary choice. Instead, respondents are asked to rate their support for the candidate using a 7-point likert scale.

In each study, respondents are randomly assigned to one of the three conditions. Because the second condition is restricted on the sensitive feature, half of the respondents are assigned to this condition in order to gain precision, while a quarter of the sample is assigned to the baseline and the third condition respectively.

<table>
<thead>
<tr>
<th>Study</th>
<th>Paired conjoint, High contrast</th>
<th>Paired conjoint, Restricted</th>
<th>Single-profile conjoint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>5 of 5 pairs contrast race</td>
<td>1 out of 5 pairs contrasts race</td>
<td>Random assignment of race</td>
</tr>
<tr>
<td>Feature: Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Levels: Black / White</td>
<td>N = 854</td>
<td>N = 1765</td>
<td>N = 874</td>
</tr>
<tr>
<td>Study 2</td>
<td>80 % chance that one candidate is Muslim</td>
<td>17 % chance that one candidate is Muslim</td>
<td>Random assignment of religion</td>
</tr>
<tr>
<td>Feature: Religion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Levels: Muslim / Protestant / Catholic / Jewish / Atheist / Other</td>
<td>N = 926</td>
<td>N = 1770</td>
<td>N = 870</td>
</tr>
</tbody>
</table>

Figure 1: Experimental conditions in Study 1 and Study 2
3.2 Study 1: U.S. Supreme Court nominees

The first study is inspired by a candidate single-profile conjoint study on support for Supreme Court nominees by Sen (2017). The design is a classic conjoint candidate design in which the researcher asks a sample of 1,650 U.S. adults to rank a number of candidates based on 8 features. While the original study used three different outcome measures on a 7-point likert scale ("Support", "Qualifications", and "Trust"), I only ask respondents to rate their support for candidates and exclude information on political leaning that was assigned to half of the respondents in the original study. In the three conjoint experiments, I used identical features, levels, wording and identical formatting. The experiments include seven features that each hold several feature levels (See details in appendix E). Most importantly, candidates’ race are assigned from a list with two levels (black and white).

In the baseline condition respondents are presented with five different pairs with maximum racial variation in each pair. That is, all five candidate pairs are Black-White or White-Black. The second condition is equivalent to the first except candidates’ race is restricted to appear only in one of the five pairs. Thus, in four pairs, both candidates are white. In the final condition, respondents are presented with a single-profile conjoint in which race is assigned randomly in order to eliminate the contrast on race that is inherent to the comparison in paired-conjoint designs.

3.3 Study 2: Immigrants seeking admission to the U.S.

Study 2 is substantively inspired by (Hainmueller and Hopkins 2015) on support for immigrants to the U.S. As was the case in Study 1, respondents were assigned to one of three variations of a candidate conjoint design that all included seven features (Gender, Age, Religious affiliation, Reason for the application, Profession, Years of experience, and prior trips to the U.S.). Most importantly here is Religion that could take on six values (Catholic, Protestant, Jewish, Muslim, Atheist or Other).

I follow roughly the same strategy as in study 1, and assign respondents to three different conjoint designs varying the focus on the sensitive level. In the baseline, a paired-conjoint design, the probability that one of the two candidates in any given pair is Muslim is high (0.8), whereas in the second condition, also a paired-conjoint, the probability that one of the two candidates is Muslim is restricted (0.17). Finally, in a single-profile conjoint, Muslim is assigned with the same probability as other religious affiliation, but the religious contrast is arguably not as prominent due to the design.

---

5In the original study, the features were chosen to approximate the information available to immigration officials which is why religion was omitted, but the authors suggest religion as a dimension for future work to explore.
4 Results

It is essential that the design variations affect respondents awareness about the sensitive item. To be sure that this was the case, the survey included a post-treatment question asking respondents to guess what they believed to be the main objective of the study from a list of eight different options. As displayed in Figure 1 a) and 1 b) respondents’ anticipation of the research objective changes drastically conditioned on design across the six experiments. In the first study, 38 percent of respondents guessed that the primary intent of the study was to examine reactions to candidates race. This underlie that race is a salient topic to which respondents pay attention. More importantly, this is 30 percentage points more than the restricted paired conjoint, and 13 percentage points more relative to the single-profile conjoint. In other words, the manipulations seemed to work as intended meaning that when downplaying focus on a sensitive dimension through design, it affect respondents inferences about the research intentions.

In Study 2, the manipulations also worked as intended although the differences between conditions are smaller. 48 percent of respondents in the baseline experiment believed that the main objective of the study was immigrants’ religious affiliation. This is 20 percentage points more compared to the restricted paired conjoint – a large difference considering that the only difference was the probability that one of the candidates in each pair was Muslim (0.8 and 0.17 respectively) compared to five other religious affiliations. In the single conjoint, 32 percent believed that religion was the main objective of the study, a significant difference to the baseline of 16 percentage points.

Figure 2: Manipulation check: Respondents perception of the research objective

---

See the questions asked in Appendix x.
The results confirm the expectations. As argued previously, we can assume that awareness about a sensitive topic increases the risk of SDB and we would expect respondents to be more favorable towards black candidates as well as the Muslim immigrant in the baseline experiment relative to the other conditions.

4.1 Comparing effects across designs

First, the three experiments in each study can be informally compared by visualizing the AMCEs from each condition. Notice that the single-profile conjoint designs rely on a different type of task (rating one profile instead of choosing between two) and a rating-based outcome measure. Hence, the effect estimates are not directly comparable with the paired designs. In both studies, the effect estimates of the sensitive feature level are strikingly similar across the three conditions. In study 1, the effect of being black is positive and significant in all of the three conditions. The effect of being black relative to the other feature levels appear slightly smaller relative to the other AMCEs in the single-profile design compared to the two paired designs. The substantial interpretation, however, is the same. In study 2, the effect of an immigrant being Muslim is negative and significant in all conditions. Again, the three experiments yield identical results. Altogether, there is no apparent difference in the core quantities of interest between the three experimental conditions.

\[\text{This is unsurprising considering evidence from previous candidate choice experiments and the fact that this experiment did not include political leaning (which can crowd out effects of demographic characteristics.)}\]

\[\text{The AMCEs in the single-profile conjoint designs are generally smaller relative to the paired designs. One way to increase the comparability between designs is by re-weighting the AMCEs of the single profile designs using general differences in the AMCEs as a weight. See appendix E}\]
Figure 3: Results from Supreme Court candidate conjoint experiments (N=3,493)
Next, I turn to a formal analysis of the two paired conjoint experiments in each study. Since they are measured on the same outcome, they are directly comparable. Hence, the effect of reducing attention to the sensitive topic can be formally tested estimating the difference-in-difference by interacting the experimental condition with the sensitive topic in each study respectively. In
study 1, this estimand is expressed as:

\[
(E[\text{choice} \mid \text{Black} \ \& \ \text{Baseline}] - E[\text{choice} \mid \text{White} \ \& \ \text{Baseline}]) \\
- (E[\text{choice} \mid \text{Black} \ \& \ \text{Restricted}] - E[\text{choice} \mid \text{White} \ \& \ \text{Restricted}])
\]

An equivalent estimand is used in study 2 where "Black" equals "Muslim" and "White" equals the reference category. A positive estimate indicate larger treatment effects in the restricted conjoint which is aligned with the expectation that SDB can be introduced by raising awareness to the sensitive feature. Figure 4 displays the differences in treatment effects between conditions. The difference between the two estimates is remarkably close to null in both studies, implying no difference in effects across designs. In study 1, the effect of being black compared to white increases the probability that a profile is chosen by 0.091 (SE = 0.016) in the baseline experiment and 0.092 (SE = 0.013) in the restricted paired conjoint.

While the effect of being Muslim is negative in both the paired design with high contrast (-0.096, SE = 0.209) and the restricted paired design (-0.073, SE = 0.017), the magnitude of the effect is fairly constant across the two paired designs. All in all, the results show no evidence that awareness about a sensitive level changes respondent behavior.
Even though respondents on average hold very different anticipations of the research objective – and have different possibilities of acting strategically – across designs, it does not change their preferences substantially.

4.2 What else could explain the null-findings?

One concern is that the increased awareness induces demand-effects that cancels out the effect of social desirability bias. A way to further bolster the results is to look at subsets of the samples that are more attentive to self-presentational concerns. Political liberals have been found to be more likely to give untruthful answers to questions regarding race (Gilens, Sniderman, and Kuklinski 1998) and immigration restrictionist policy questions (Janus 2010). Moreover, previous studies have found that high self-monitors are more likely to give appropriate answers to sensitive questions.

In both studies, I reestimate the difference-in-differences looking at respondents that identify as liberal.\(^9\) Moreover, following Berinsky and Lavine (2011),

\(^9\)respondents with a score >6 on a 0-10 scale ranging from "Very conservative" to "Very liberal".
study 2 included three items from the self-monitoring scale that was also used by Hainmueller and Hopkins (2015). As shown in Figure B in the supplementary material, the difference-in-differences from the liberal subset is very close to null (-0.014 SE = 0.031; 0.0007 SE = 0.049 respectively). The same is true for the subset of high self-monitors in study 2 (.020 SE = 0.043).

Another concern is, that the "treatment" in the paired design with high contrast was not assigned before the experiments, but rather works as an integrated part of the design. Hence, we might believe that the sensitive dimension became increasingly obvious to respondents as they worked their way through the five conjoint pairs. In other words, respondents assigned to the paired design with high contrast were more aware about the sensitive feature when they where asked to choose between a black and a white candidate for the third, fourth and fifth time. Consistent with the expectations, we would expect the repetition of choosing between two candidates that are contrasted on a sensitive dimension to increase SDB – and therefore change results towards more politically correct answers through the survey. In order to test this, I compare estimates in the baseline experiments from pair 1-5 respectively. The change in effect sizes as respondents answer the five pairs in the baseline experiments are inconsequential and does not support the notion that respondents change preferences as the contrast on a sensitive dimension is repeated (Figure C in supplementary material).

5 Conclusion and discussion

Conjoint designs are often claimed to limit concerns over social desirability bias: that research subjects respond in ways that are socially desirable. It is argued that due to the large number of features, respondents cannot infer the main intent of the experiment and can always justify any given answer. However, to what extent this is the case might depend on the specific research design. The present study tests this by randomly assigning respondents to different types of conjoint designs.

The results provide evidence that the design of conjoint experiments have an effect on respondents’ inferences about the main objective: they pay more attention to a sensitive feature in the paired conjoint with frequent contrast. However, the core quantities of interest are remarkably stable across designs.

There are several implications of these results. First, respondents do not change behavior when asked to choose their preferred candidate across two paired conjoint designs assigning a sensitive level with different probability weights. This is true even though the paired conjoint design contrasts a sensitive feature repeatedly across pairs. The stability of results also goes against

10The following questions are used: “When you’re with other people, how often do you put on a show to impress or entertain them?” Response categories: Always, Most of the time, About half the time, Once in a while, Never. “How good or bad of an actor would you be?” Response categories: ‘Excellent’, ‘Good’, ‘Fair’, ‘Poor’, ‘Very poor’. “When you are in a group of people, how often are you the center of attention?” Response categories: ‘Always’, ‘Most of the time’, ‘About half the time’, ‘Once in a while’, ‘Never’. 
recent suggestions that paired-conjoint designs makes it easier for respondents to act strategically.

Secondly, while this study cannot rule out SDB in conjoint experiments, it is reassuring that the paired and the single-profile conjoint designs give the same results. This suggests that results from these designs are not an artifact of the design, but rather mirror respondents’ priorities. Thirdly, and consequently, the results suggest that there is no immediate reason to choose a design that is sub-optimal in order to “disguise” sensitive topics. Researchers might choose a design that restrict probabilities of certain features (which will also reduce statistical power) or use a single-profile (although a paired design might be preferable) to avoid SDB. While there can be other arguments for these designs, this study finds no reason to do so because of concerns of SDB. Beyond conjoint designs specifically, it supports recent evidence by Mummolo and Peterson (2018) and De Quidt, Haushofer, and Roth (2017) that experimental researchers should not be too concerned with respondents’ awareness to research intentions.

References


Horinuki, Yusaku, Daniel M Smith, and Teppei Yamamoto (2017). “Identifying Voter Preferences for Politicians’ Personal Attributes: A Conjoint Experiment in Japan”. In:


Ono, Yoshikuni and Masahiro Yamada (2016). “Do Voters Prefer Gender Stereotypic Candidates?: Evidence from a Conjoint Survey Experiment in Japan”. In:
Schwarz, Susanne, William Hunt, and Alexander Coppock (2018). “What Have We Learned About Gender From Candidate Choice Experiments? A Meta-analysis of 30 Factorial Survey Experiments”. In:
Supplementary material: Social desirability bias in candidate conjoint experiments: What is the optimal design when studying sensitive topics?
Appendix A

Recruitment and Response Rate

The experiments were implemented in Qualtrics and fielded at Amazon’s Mechanical Turk. The sampling took place between August 8 and August 30. The sampling design was a random sampling using the build-in randomize option in Qualtrics. Only respondents who answered the last question (the manipulation check) are included in the final sample.

The respondents were presented with a paired design or a single-profile design. Screenshots of a paired conjoint design and a single-profile conjoint design are shown in Figure 6 and 7.

![Figure 6: Example of discrete choice conjoint](image)

![Figure 7: Example of rating-based conjoint](image)
Appendix B

Table E1 give details on the features and feature levels used to generate the profiles in Study 1

<table>
<thead>
<tr>
<th>Feature</th>
<th>Level</th>
<th>Feature</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male, Female</td>
<td>Gender</td>
<td>Male, Female</td>
</tr>
<tr>
<td>Age</td>
<td>25-75 (Continuous)</td>
<td>Age</td>
<td>25-75 (Continuous)</td>
</tr>
<tr>
<td>Race</td>
<td>White, Black</td>
<td>Religion</td>
<td>Atheist, Protestant, Jewish, Muslim, Catholic, Other</td>
</tr>
<tr>
<td>Education</td>
<td>Yale Law School, Florida State Uni., Albany Law School</td>
<td>Reason for application</td>
<td>Seek better job, Reunite with family members, Escape religious/political persecution</td>
</tr>
<tr>
<td>Religion</td>
<td>Mormon, Mainline Protestant, Jewish, Evangelical Protestant, Catholic</td>
<td>Profession</td>
<td>Doctor, Nurse, Teacher, Waiter, Construction worker, Computer programmer</td>
</tr>
<tr>
<td>Clerksip experience</td>
<td>Did not serve as law clerk, did serve as law clerk</td>
<td>Working experience</td>
<td>None, 1-2 years, 3-5 years, More than 5 years</td>
</tr>
<tr>
<td>Previous work experience</td>
<td>Elected politician, law professor, lawyer in private practice, non-profit lawyer, public defender</td>
<td>Prior trips to the U.S.</td>
<td>Never been to the U.S., Spent six months with family, Visited once without legal authorization, Visited once on tourist visa, Visited many times on tourist visa</td>
</tr>
</tbody>
</table>

Figure 8: Features and feature levels included in the conjoint experiments

Appendix C

One way to further bolster the results is to look at a subset of the sample that are more attentive to self-presentational concerns. First, I reestimate the difference-in-differences in both studies comparing respondents across score on a 0-10 scale ranging from “Very conservative” to “Very liberal”. Political liberals have been found to be more likely to give untruthful answers to questions regarding race (Gilens, Sniderman, and Kuklinski 1998) and immigration restrictionist policy questions (Janus 2010), and we would therefore expect to see stronger SDB bias among liberals. Secondly, previous studies have found that high self-monitors are more likely to give appropriate answers to sensitive questions. Following Berinsky and Lavine (2011), I used three items from the self-monitoring scale that was also used by Hainmueller and Hopkins (2015). 11

11The following questions are used: “When you’re with other people, how often do you put on a show to impress or entertain them?” Response categories: Always, Most of the time, About half the time, Once in a while, Never. “How good or bad of an actor would you be?” Response categories: Excellent, Good, Fair, Poor, Very poor. “When you are in a group of people, how often are you the center of attention?” Response categories: Always, Most of the time, About half the time, Once in a while, Never.
Appendix D

The treatment was not assigned before the experiments, but rather is embedded in the design. Hence, respondents in the paired design with high contrast may have been more aware about the sensitive feature when they where asked to choose between a black and a white candidate for the third, fourth and fifth time. To test this, I compare estimates in the baseline experiments from pair 1, 2, 3, 4 and 5 respectively. As shown in Figure C, the change in effect sizes as respondents answer the five pairs in the baseline experiments are inconsequential and does not support the notion that respondents change preferences as the contrast on a sensitive dimension is repeated.
Appendix E

Re-weighted results The single-profile design and the paired designs are not directly comparable. First of all, the tasks that respondents were asked to solve differ: in the paired designs they are presented with two profiles while in the single-profile, they only see one at a time. Moreover, the outcome (either forced choice or rating) is different as well. Second of all, the AMCEs in the single-profile conjoint designs are generally smaller compared to the paired designs. This makes the direct comparison of the effect estimate of the sensitive feature across design problematic. One way to solve this is to re-weight the AMCEs of the single profile designs. In general, we can comprehend the magnitude of a given AMCE is by comparing the effect estimate of that given feature relative to effect estimates of other features in the same conjoint design. Hence, a way to increase comparability between the single-profile design and the paired designs is to re-weight the effect of the sensitive levels by using the average between-design difference in other AMCEs. In other words, I find the average difference in effect estimates of all other features between the single-profile and the paired designs. On that basis, the estimate of the sensitive feature is re-weighted.

In study 1, all other features than race are on average larger by a factor 1.33 relative to the effects in the single-profile conjoint. Hence, re-weighting the effect of race in the single-profile gives a new, larger effect estimate (.063), which is closer to the paired designs. In Study 2, effects of other features are on average 2.4 times larger than in the paired designs. Hence, re-weighting the effect of being Muslim increases the effect from .07 to .16, meaning that the negative effect of being Muslim becomes larger than the estimate in both the paired designs. Column 3 give the unweighted effect estimate, while column 4
give the re-weighted effect estimate of the single-profile experiment.