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 experimental designs is that conjoint experiments have the potential to mitigate social desirability bias. This is based on the notion that since respondents are presented with numerous features, the research objective is concealed and respondents can find multiple justifications for inappropriate answers. However, if and to what extent this is true may depend on specific design choices – a concern that has received little empirical attention. In this paper, I randomly assign respondents to six different conjoint designs in order to manipulate respondents’ awareness to sensitive features and their possibilities of acting strategically when responding. The results show that while design variations have substantial effects on respondents’ awareness to sensitive features, there are no detectable effects on respondents’ behavior. The evidence suggests that the fear of social desirability bias should not induce researchers to use conjoint designs that are otherwise sub-optimal.

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 in political elections (Teele, Kalla, and Rosenbluth 2018), how information on candidates party affiliation moderate voters’ preferences for candidates’ demographic characteristics (Kirkland and Coppock 2017), and attitudes towards immigrants (Hainmueller and Hopkins 2015). Conjoint experiments enable researchers to examine multidimensional preferences over candidates and allow for comparing the relative importance of a large number of 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; Teele, Kalla, and Rosenbluth 2018). This is an essential quality, especially considering that these designs are often used to gauge respondents’ reactions to sensitive dimensions such as candidates’ race or sex. In fact, the usefulness of conjoint analysis 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 a concern when examining sensitive topics.

That conjoint designs can mitigate SDB builds on the notion that since respondents are presented with numerous features, they can not anticipate the research objective.1 Moreover, respondents can always find multiple justifications for any given choice. However, the extent to which this is true may be largely conditioned on a number of specific design choices researchers have to make when conducting conjoint designs. For example, there is a fundamental difference in terms of whether the hypothetical candidate profiles are shown in pairs or not, and outcomes can be based on either a discrete choice, a rating of the candidates, or a combination of the two. Moreover, the number of features varies with studies displaying a few up to a dozen different features.2 Finally, 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 with different probability weights. These design differences may greatly affect (1) respondents’ inferences about the primary research objective, (2) their possibilities for acting strategically when responding, and (3) the degree to which respondents can justify inappropriate answers over repeated tasks. In other words, the degree to which SDB is a concern in conjoint designs may depend on specific design choices.

In this study, I seek to identify if design decisions can introduce SDB in candidate conjoint studies and what researchers can do to avoid it. To do so, I

---

1This is not a new idea. Respondents in survey experiments are aware that they are being evaluated by an unseen but omnipresent scholar (Berinsky 2004) and therefore experimentalists often attempt to avoid revealing their intentions or include strategies to deceive survey participants about the research agenda (McDermott 2002; Berinsky, Huber, and Lenz 2012).

2Features can for example be Age/Party affiliation/Race whereas feature levels are the values each feature can take, e.g. Male/female.
run a total of six candidate conjoint experiments that all include sensitive topics. The studies rely on convenience-samples and were conducted at Amazon’s Mechanical Turk (N = 7,059). Respondents are randomly assigned to conjoint designs intended to either downplay or amplify their attention to sensitive dimensions and their possibilities of acting strategically. Specifically, I examine two design components. First, if some sensitive feature levels are repeatedly contrasted across a number of conjoint pairs – for example a black vs. a white candidate – we may fear that respondents infer that the sensitive feature (in this case ‘race’) is the primary research objective. Moreover, if respondents are asked repeatedly to choose between candidates that differ on a sensitive feature, it arguably reduces their possibilities of justifying socially inappropriate answers. As a consequence, the risk that respondents give dishonest answers increase. Secondly, paired-designs may raise respondents awareness to sensitive features due to the within-subject structure, for example, by displaying a black and a white candidate at the same time. Moreover, since respondents are presented with control and treatment simultaneously it is easier for them act strategically (Mummolo and Peterson 2018). I therefore compare results from paired designs with a single-profile design in which respondents only see one hypothetical candidate at a time.

The results show that variations in conjoint design have large effects on the way respondents perceive the experiment. 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. 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 feature and in no instance did the experiments produce significantly distinguishable effects. The implications for the design of conjoint experiments are discussed.

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 (Janns

---

3This 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 Quitt, Haushefer, and Roth 2017). However, instead of raising awareness to the research objective by providing respondents with explicit information, the present project seek to manipulate awareness to sensitive dimensions through design.
SDB thus masks the effect of in-group favoritism and lead to a misrepresentation of preferences. Evidence on SDB in survey studies in general suggests that it is a valid concern. This is explored in studies that word questions in less threatening ways (Kuklinski et al. 1997), change the interview setting (Krysan and Couper 2003), or compare list experiments with direct questions (Janus 2010; Gilens, Sniderman, and Kuklinski 1998). Finally, some studies examine 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.\footnote{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 which I test in the final part of the paper.}

### 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, due to 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 in evaluations of political candidates, 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 in order to mitigate SDB is not a new idea. Previous survey 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).

The notion that conjoint designs have features that should mitigate concerns over SDB appear sound, but there is little empirical evidence to support it. Some studies relying on candidate conjoint studies implement various tests
in order to examine potential SDB. For example, Bansak, Hainmueller, and Hangartner (2016) find that results are stable for respondents with different levels of empathy, building on the idea that empathy and social desirability scales correlate. Hainmueller and Hopkins (2015) come to the same conclusion after re-estimating 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.

So why should SDB be a concern? Results from several candidate conjoint experiments on sensitive dimensions seem add odds with what we know from field experiments or observational studies. For example, a number of political candidate conjoint designs find no effects – or even positive effects – of being black compared to white (Carnes and Lupu 2016; Kirkland and Coppock 2017).\footnote{Carnes and Lupu (2016) conduct a conjoint experiment in which they manipulate candidates’ race using two levels (white and 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 (although this is not significant).} Moreover, despite considerable study-to-study variation, a majority of political candidate conjoint experiments that include gender show a net preference for women (Schwarz, Hunt, and Coppock 2018). While this can of course reflect respondents’ true preferences, we may worry that the experiments overestimate support for female and black candidates due to SDB. 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). Hence, it is worthwhile to consider the way design decisions impact respondent behavior.

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.\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 not to replicate these studies, but rather to determine whether treatment effects vary across design, it is not an issue.} 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 feature that is known to be influenced by social desirability pressures. Study 1 seeks 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 (Krupnikov, Piston, and Bauer 2016; Berinsky and Lavine 2012). Study 2 seeks to unravel support for immigrants seeking admission to the US. Religious affiliation, more specifically being Muslim, serve as a sensitive feature level. 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.\footnote{The Muslim ban set off a fury of protests across U.S. cities and airports, garnering tremendous media attention and discussion (Collingwood, Lajevardi, and Oskooi 2018)}

I seek to manipulate respondents’ awareness to the sensitive feature in two ways. First, I manipulate the probability weights of the levels of the sensitive feature across conditions. Thus, one condition, the high-contrast design, is a paired-conjoint in which respondents are presented with five different candidate pairs with each or most pairs displaying a contrast on the sensitive feature (for example, a black vs. a white candidate). Arguably, the repeated contrast increase respondents’ awareness to the sensitive feature. Moreover, the frequent contrast makes it harder for respondents to defend an inappropriate answer if they have to repeat it across five conjoint pairs. We can expect SDB to amplify in this condition. In the second condition, the restricted paired design, the sensitive feature is contrasted less frequently. Thus, the restriction serve to "mask" the sensitive feature from respondents by design.

Secondly, I test the importance of the within-subject structure by including a single-profile design as a third condition. While respondents in the paired conjoint designs observe both treatment and control at the same time, the single-profile conjoint displays either control or treatment which arguably make the sensitive feature less noticeable. Moreover, the single-profile design makes it harder for respondents to act strategically compared to the within-subject structure of the paired design (Mummolo and Peterson 2018).

In each study, respondents are randomly assigned to one of the three conditions. Because the second condition is restricted on the sensitive feature which reduces statistical power, half of the respondents are assigned to this condition in order to gain precision, while a quarter of the sample is assigned to the high-contrast design and the single-profile design respectively.

\subsection*{3.2 Study 1: U.S. Supreme Court nominees}

The first study is inspired by a 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 hypothetical candidates. 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. Respondents are randomly assigned to one of the three conjoint experiments that are otherwise identical in terms of features, levels, wording and formatting. The experiments include six 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 or white).

In the high-contrast condition respondents are presented with five different pairs in which each pair has variation in race. That is, all five candidate pairs appear as Black vs. White or White vs. 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 order to eliminate the contrast on race that is inherent to the comparison in paired-conjoint designs. In this condition the candidates’ race is assigned randomly.

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

Study 2 is substantively inspired by Hainmueller and Hopkins (2015) and examines support for immigrants applying for admission to the U.S. As is the case in Study 1, respondents are assigned to one of three variations of a candidate conjoint design that all include seven features (See details in appendix E). Most importantly here is Religion that can take on six levels (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 high-contrast design, the probability that one of the two candidates in any given pair is Muslim is high (80 per cent of all pairs), whereas in the second condition the probability that one of the two candidates is Muslim is restricted (17 per cent of all pairs). Finally, in a single-profile conjoint, religious affiliation is drawn randomly, but as was the case in study 1, the religious contrast is arguably not as prominent due to the non-paired design.

---

8In 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

4.1 Can respondents infer research intentions?

A first-order concern is whether the design variations in fact have an effect on respondents' awareness to the sensitive feature. To check if this is the case, the survey included a post-treatment question asking respondents to choose from a list of eight different options what they believed to be the main objective of the study. As displayed in Figure 2 respondents' anticipation of the research objective changes drastically across design conditions. In Study 1, 38 percent of respondents in the high-contrast condition guessed that the primary intent of the study was to examine reactions to candidates race. This is 30 percentage points more compared to the restricted paired conjoint, and 13 percentage points more relative to the single conjoint. In other words, researchers can greatly influence respondents' inferences about research intentions through the design. When downplaying a sensitive dimension, respondents are on average less likely to guess that the sensitive feature is important.\(^9\)

\(^9\)More generally the large differences across design indicate that Mturkers pay attention to the tasks.

<table>
<thead>
<tr>
<th>Study 1</th>
<th>Paired conjoint, High contrast</th>
<th>Paired conjoint, Restricted</th>
<th>Single-profile conjoint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature: Race</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>Levels: Black / White</td>
<td>N = 854</td>
<td>N = 1765</td>
<td>N = 874</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study 2</th>
<th>Paired conjoint, High contrast</th>
<th>Paired conjoint, Restricted</th>
<th>Single-profile conjoint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature: Religion</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>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
Assuming that respondents give more desirable answers when they anticipate that a sensitive feature is the main research objective, we would expect respondents to give different answers across conditions. More specifically, respondents should be more favorable to the black political candidates as well as the Muslim immigrants in the high-contrast design relative to the restricted paired design and the single-profile design.

4.2 Does design variations affect respondent behavior?

Before formally testing the differences across designs, the AMCEs are visualized in Figure 3. Notice that the single-profile conjoint designs rely on a different type of task (evaluating one profile at a time instead of choosing between two) and a rating-based outcome measure. Hence, the effect estimates are not immediately comparable with the paired designs. Yet, the effect estimates of the sensitive feature levels are strikingly similar across the three conditions in both studies. In study 1, the effect of being black is positive and significant in all of the three conditions. In study 2, the effect of being Muslim is negative and significant in all conditions. Altogether, there is no apparent difference in the core quantities of interest between the three experimental conditions.

---

10This 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.
Figure 3: Results from Supreme Court candidate conjoint experiments (N=3,493)
Next, I turn to a formal analysis of the design effects. The outcome of interest is the differences in effects of the sensitive feature levels across designs. First, the two paired conjoint experiments in each study are compared. They rely on the same outcome and are therefore directly comparable. Hence, the effect of reducing attention to the the sensitive topic can be formally tested.
estimating a difference-in-difference. In other words, I interact a design dummy variable (high-contrast = 0 / restricted = 1) with the sensitive topic in each study respectively. A positive estimate indicates that respondents give more desirable answers in the high-contrast design which aligns with the expectation that SDB can be introduced by raising awareness to the sensitive feature. As shown in Figure 5 (a) the difference in the effects of the sensitive feature between designs is remarkably close to null in both studies.

In Figure 5 (b) I follow the same strategy in order to compare the high-contrast paired design and the single-profile design. However, since the experiments rely on different outcomes the comparison is not as straightforward. Since the AMCEs are consistently smaller in the single-profile design, the size of the effect of a candidate being Black or Muslim is naturally smaller compared to the paired designs. I account for this by weighting the AMCEs in the single-profile design using the relative difference of all AMCEs between the single-profile and the paired designs as a weight (See details in appendix x). As seen in Figure 5 (b), there are no substantial differences between the single-profile and the high-contrast paired designs either.

Altogether, the results show that even when respondents anticipate a sensitive feature as important and at the same time have optimal conditions for tailoring their answers, it does not change their behavior.

11 For example, the estimand comparing the two paired design is expressed as:

\[
(E\{choice \mid Black & high-contrast\} - E\{choice \mid White & Highcontrast\}) - (E\{choice \mid Black & Restricted\} - E\{choice \mid White & Restricted\})
\]

An equivalent estimand is used in study 2 where "Black" equals "Muslim" and "White" equals the reference category.

12 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 high-contrast design and 0.092 (SE = 0.013) in the restricted paired conjoint. The effect of being Muslim is negative in both the high-contrast design with a coefficient of -0.096 (SE = 0.209) and -0.073 (SE = 0.017) in the restricted paired design.

13 Specifically, the paired designs give AMCEs that are on average larger by a factor 1.87. The notion of re-weighting the results rely on a core feature of conjoint experiments: that AMCE of any feature level can be gauged relative to estimates of other feature levels. By weighting the results, all AMCEs in the single-profile change with a constant, but the relative difference between each AMCE remain the same.
4.3 What else could explain the null-findings?

The results are good news to researchers conducting conjoint experiments: we should not be too concerned with implementing designs that, at least in in theory, increase the risk of SDB. In this section I test alternative explanations for the null-findings.

One concern is that the increased awareness to a sensitive feature also introduces demand-effects that cancels out SDB. Because respondents may anticipate that the researcher expect to find a bias against black or Muslim candidates, they might "help" the researcher confirm the hypothesis, which would bias the effect in the opposite direction than SDB. Hence, to further bolster the results, I look at subsets of the samples that are more likely to be attentive to self-presentational concerns – and thus where we would expect SDB to be most pronounced. 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.\textsuperscript{14}

Moreover, following Berinsky and Lavine (2011), study 2 included three items

\textsuperscript{14}Respondents with a score >6 on a 0-10 scale ranging from "Very conservative" to "Very liberal".
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 and 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 is embedded in the design. Hence, it is possible 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 could 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. In that case, results should change towards more politically correct answers towards the end of the conjoint. To test this, I compare estimates in the high-contrast designs from pair 1-5 respectively. The change in effect sizes as respondents answer the five pairs in the high-contrast designs 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 moreover, they can always justify any given answer. However, to what extent this is true depends on specific research design decisions. The present study tests this by randomly assigning respondents to different variations 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 than they do in designs where the contrast is less obvious. However, the core quantities of interest are remarkably stable across designs.

There are several implications of the results. First, while this study cannot rule out that SDB is an issue in conjoint experiments, it is reassuring that different types of conjoint designs give the same results. This suggests that results are not an artifact of the design, and that we do in fact measure respondents’ priorities. Second, the stability of results also goes against recent suggestions that paired-conjoint designs makes it easier for respondents to act

---

15 The 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'.
strategically. There is no evidence that respondents act strategically when presented with a within-subject design. Thirdly, and consequently, there is no immediate reason to choose a design that is otherwise sub-optimal in order to disguise sensitive topics. There is no immediate reason for restricting probabilities of certain feature levels (which also reduces statistical power) or for using a single-profile design (when a paired design is preferable) to avoid SDB. Finally, beyond conjoint designs specifically, the present study supports recent evidence by Mummolo and Peterson (2018) and De Quidt, Haushofer, and Roth (2017) that researchers should not be too concerned with respondents’ awareness to research intentions in survey experiments.

References


Hainmueller, Jens, Dominik Hangartner, and Teppi Yamamoto (2015). “Validating vignette and conjoint survey experiments against real-world behav-


Horiuchi, 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

Figure 7: Example of rating-based conjoint
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>Clerkship 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-presentation 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 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). The 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 high-contrast designs 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 high-contrast design are inconsequential and does not support the notion that respondents change preferences as the contrast on a sensitive dimension is repeated.

Figure 9: Dif-in-dif liberal respondents
### 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.