The Number of Choice Tasks and Survey Satisficing in Conjoint Experiments

Kirk Bansak[1], Jens Hainmueller[2], Daniel J. Hopkins[3] and Teppei Yamamoto[4]

[1] *PhD Candidate, Department of Political Science, 616 Serra Street Encina Hall West, Room 100, Stanford, CA 94305-6044, USA. Email: kbansak@stanford.edu*
[2] *Professor, Department of Political Science and Graduate School of Business, 616 Serra Street Encina Hall West, Room 100, Stanford, CA 94305-6044, USA. Email: jhain@stanford.edu*
[3] *Associate Professor, Department of Political Science, University of Pennsylvania, 207 S. 37th Street, Philadelphia, PA 19104, USA. Email: danhop@sas.upenn.edu*
[4] *Associate Professor, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. Email: teppei@mit.edu, URL: http://web.mit.edu/teppei/www*

Abstract

In recent years, political and social scientists have made increasing use of conjoint survey designs to study decision-making. Here, we study a consequential question which researchers confront when implementing conjoint designs: How many choice tasks can respondents perform before survey satisficing degrades response quality? To answer the question, we run a set of experiments where respondents are asked to complete as many as 30 conjoint tasks. Experiments conducted through Amazon's Mechanical Turk and Survey Sampling International demonstrate the surprising robustness of conjoint designs, as there are detectable but quite limited increases in survey satisficing as the number of tasks increases. Our evidence suggests that in similar study contexts researchers can assign dozens of tasks without substantial declines in response quality.

Keywords: survey experiments, response bias, survey design

1 Introduction

First introduced in the 1970s (Green and Rao 1971; Jasso and Rossi 1977), conjoint experiments ask survey respondents to rate or rank profiles which vary across multiple dimensions. This research design has critical strengths: It allows researchers to make causal inferences about a variety of potentially relevant attributes simultaneously, and so to compare the treatment effects of various attributes (Hainmueller, Hopkins, and Yamamoto 2014). Conjoint designs also mirror many real-world choices in which people must evaluate bundles of attributes, which can greatly enhance their external validity (Hainmueller, Hangartner, and Yamamoto 2015). These characteristics, together with the increasing number of surveys administered via computers, have led to a surge in the use of conjoint designs in political science. Conjoint designs are now being used to answer far-ranging questions, including those about where people choose to live, whom they wish to admit to their countries, and which political candidates they support.[1]

While research proceeds on the statistical properties of conjoint designs (Raghavarao, Wiley, and Chitturi 2011; Hainmueller, Hopkins, and Yamamoto 2014; Egami and Imai 2015; Acharya, Blackwell, and Sen 2016), there has been little attention on how to optimize conjoint survey

Political Analysis (2018)
vol. 26:112–119
DOI: 10.1017/pan.2017.40

Corresponding author
Teppei Yamamoto

Edited by
Lonna Atkeson

© The Author(s) 2018. Published by Cambridge University Press on behalf of the Society for Political Methodology.

Authors' note: We thank the associate editor, two anonymous referees, Katrin Auspurg, Adam Berinsky, Thomas Hinz and the conference participants at the MPSA 2017 Annual Meeting, PolMeth XXXIV, and the University of Zurich for their helpful comments and suggestions. Replication materials for this article are available on the *Political Analysis* Dataverse as Bansak *et al.* (2017b).

1 For example, see Franchino and Zucchini (2015), Abrajano, Elmendorf, and Quinn (2015), Carnes and Lupu (2015), Hainmueller and Hopkins (2015), Horiuchi, Smith, and Yamamoto (2015), Bansak, Hainmueller, and Hangartner (2016), Bechtel, Genovese, and Scheve (2017), Mummolo and Nall (2017), Wright, Levy, and Citrin (2016).

https://doi.org/10.1017/pan.2017.40 Published online by Cambridge University Press

designs given well-known challenges in survey research. Here, we integrate research on survey design with work on conjoint experiments to examine a central design question facing those fielding conjoint experiments: How many conjoint tasks can respondents perform with the needed levels of attention?

Underlying the question is the threat of survey satisficing. Research on survey taking indicates that as survey tasks become more onerous, respondents become increasingly likely to satisfice, meaning that they adapt by using cognitive shortcuts which can degrade response quality (Krosnick 1999).[2] Satisficing respondents are more likely to rush through surveys, ignore or skip instructions, choose response options because of their placement, and use other effort-saving heuristics (Berinsky, Margolis, and Sances 2014).

In this paper, we conduct a series of conjoint survey experiments to empirically examine the degree of satisficing when respondents are faced with a large number of choice tasks. In our experiments, we ask respondents to complete many more tasks than in a typical conjoint study and estimate the degree of degradation in response quality over those tasks. Specifically, respondents are asked to evaluate as many as 30 conjoint tables, where the tables are comprised of two core attributes that are included for all respondents and two to 18 additional attributes that are randomly assigned for each respondent. We find that conjoint designs are remarkably robust as a tool for eliciting preferences about multidimensional objects. Using samples from two common online sources of survey respondents—Amazon's Mechanical Turk (MT) and Survey Sampling International (SSI)—we see no significant decline in the core attributes' effects as the number of tasks increases.

2 Problem: Satisficing in Conjoint Experiments

Conjoint experiments are one variant of survey research, meaning that many insights about survey design generally should apply. However, the growing research on conjoint surveys has not yet incorporated the insights of the highly developed literature on survey measurement (e.g. Groves *et al.* 2011). Here, we focus on the issue of survey satisficing and how it might undermine conjoint experiments.

A key element of conjoint designs that has the potential to increase satisficing beyond acceptable levels is the number of discrete evaluation tasks requested of respondents. Should respondents perform just one evaluation in a given survey, or should they be asked to perform 5, 10, or even 50? Conjoint experiments typically require respondents to complete the same task repeatedly. In fact, in traditional conjoint designs, respondents are often asked to evaluate the entire set of possible combinations from an orthogonalized array of attribute levels, a number which can easily grow above 50 (Raghavarao, Wiley, and Chitturi 2011). While fully randomized designs allow researchers more discretion in choosing the number of tasks, researchers still have an incentive to assign numerous tasks so as to increase their statistical power.

However, research on survey response indicates that satisficing is likely to be a function of the total survey length. For example, Galesic and Bosnjak (2009) find that when answering questions placed later in a questionnaire, respondents take less time and provide more uniform answers. Similarly, respondents are more likely to give the same response to blocks of questions when those questions are found later in a questionnaire (Herzog and Bachman 1981), another indication of increased satisficing. Findings like these fuel the "longstanding view that long questionnaires or interviews should be avoided," even as others contend that the evidence underpinning that view is weaker than many suspect (De Vaus 2014, p. 111). Still, concerns about questionnaire length may be particularly acute when choosing the number of conjoint tasks, as fatigue may set in more rapidly when performing the same task repeatedly.

2 The term "satisficing" has various meanings within different research literatures. Here, we use the term exclusively as an abbreviated form of "survey satisficing" (see also Kahneman 2003).

https://doi.org/10.1017/pan.2017.40 Published online by Cambridge University Press

In short, researchers have good reason to expect that conjoint designs with a large number of tasks could produce significant levels of survey satisficing, but to date, there has been little empirical evidence as to the severity of this problem. Researchers are often tempted to ask respondents to complete many conjoint tasks in a single survey so as to maximize their statistical power. But this temptation carries risks, as researchers may well induce suboptimal levels of survey satisficing. Below, we provide an empirical assessment of this trade-off.

3 Empirical Evidence on the Number of Choice Tasks and Satisficing

Our goal is to investigate whether asking respondents to complete many repeated conjoint tasks will degrade their response quality due to satisficing, and if so when the degradation tends to kick in. In this section, we report the result of the six conjoint experiments we conducted for this purpose.

3.1 Design and methodology

The main portion of our study—the first five of the six experiments—was fielded on a total of 4,921 respondents we recruited via MT for payments of $1.25.[3] Our last, sixth survey was conducted on 1,613 respondents from SSI to confirm that key findings were not specific to MT respondents. These surveys occurred between February and May, 2015. While both use opt-in survey samples, MT draws from a small, highly experienced population (e.g. Stewart *et al.* 2015). As a result, by conducting our study via both MT and SSI, we can observe the role of fatigue for populations with different levels of survey-taking experience.[4]

After a few introductory demographic questions about their own education, partisanship, and ideology, we told respondents: "This study is about voting and about your views on potential candidates for President. We are going to present pairs of hypothetical presidential candidates in the United States. For each pair, please indicate which of the two candidates you would prefer to see as President." One example of the resulting conjoint task is available in Figure 1. We developed a set of twenty possible attributes that could define U.S. presidential candidates, including everything from their education, income, religion, and political partisanship to their positions on key issues (e.g. gay marriage, health care, abortion) and personal facts such as their favorite professional sport and car. The full list of attributes is provided in Table A.1 in the supplementary materials.

We employed several randomizations, some of which we report elsewhere. For one thing, we randomly varied the total number of attributes presented to respondents. Specifically, each respondent was randomly assigned to 4, 5, 6, 7, 8, 10, 15, or 20 attributes. Of those, the two "core" attributes—candidates' education and partisanship—were always included in each respondent's table regardless of their assigned number of attributes, and the rest were randomly drawn from the master list of 20 attributes. Once a specific number and set of attributes was assigned to a respondent it was fixed for the duration of her survey. We also randomized the attributes' order within the conjoint table and then fixed that order across tasks for each respondent. For example, if a respondent saw the candidate's income at the top of the conjoint table, it remained in that position for the duration of her tasks.[5]

3 The numbers of respondents for these five experiments were 605, 674, 725, 1,340, and 1,577, in the chronological order they were fielded.

4 See Bansak *et al.* (2017b) for replication materials.

5 This is an example of the multiply randomized conjoint design proposed by Hainmueller, Hopkins, and Yamamoto (2015). We also randomly varied several other elements of these experiments for the separate analyses reported in that paper. Specifically, we randomized the two core attributes to appear in the middle of the table or at the bottom (first and third experiments) and at the top or at the bottom (second, fourth and fifth experiments). Note that all of these six experiments used 30 tasks per respondent, so any design element that differed across the experiments is balanced across the 30 tasks.

https://doi.org/10.1017/pan.2017.40 Published online by Cambridge University Press

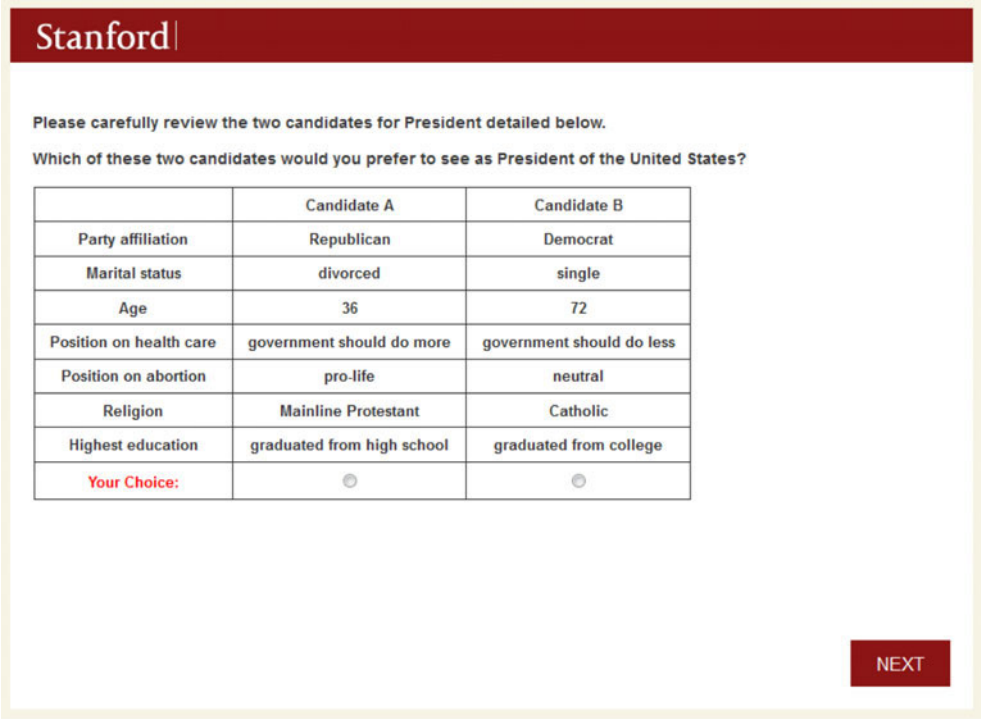

Figure 1. An example choice task from the study. Respondents are asked to assess two hypothetical presidential candidates.

Most importantly for our purposes here, we asked respondents to complete 30 of these conjoint tasks, which is much more than typical recent applications of randomized conjoint analysis. The purpose of this design choice, of course, was to study how response quality might change as respondents went through numerous screens of conjoint tables (which also consisted of a large number of attributes for some).

As the number of tasks increases, do respondents adapt by being less discerning in their choices? Our expectation is that any increased survey satisficing will induce respondents to pay less attention to the task, and so will attenuate the predictive power of the core attributes. We employ two metrics to measure the predictive power of the attributes. First, we estimate the Average Marginal Component Effects (AMCEs) of the two core attributes and compare the estimates across tasks.

Second, we calculate the coefficient of determination (i.e. R^2) from the regression of conjoint responses on the core attributes,[6] and compare those R^2s across tasks. Because the R^2 is a function of the regression-based estimates of the AMCEs under the fully randomized conjoint design, any changes in the R^2 across tasks can be attributed to changes in satisficing. In other words, the R^2 can be interpreted as a summary measure of the explanatory power of the two core attributes combined, and its change as the overall variation in satisficing.

3.2 Results

We first present results from five surveys on MT respondents. Figure 2 shows the estimated AMCEs for the two core attributes that were always included in the conjoint table—education and party affiliation—across the number of completed tasks along with their 95% confidence intervals clustered by respondent. Remarkably, the results suggest a surprising degree of robustness over

6 Specifically, we create indicator variables for all levels of each of the core attributes except for a reference level and regress the outcome on all the indicators.

https://doi.org/10.1017/pan.2017.40 Published online by Cambridge University Press

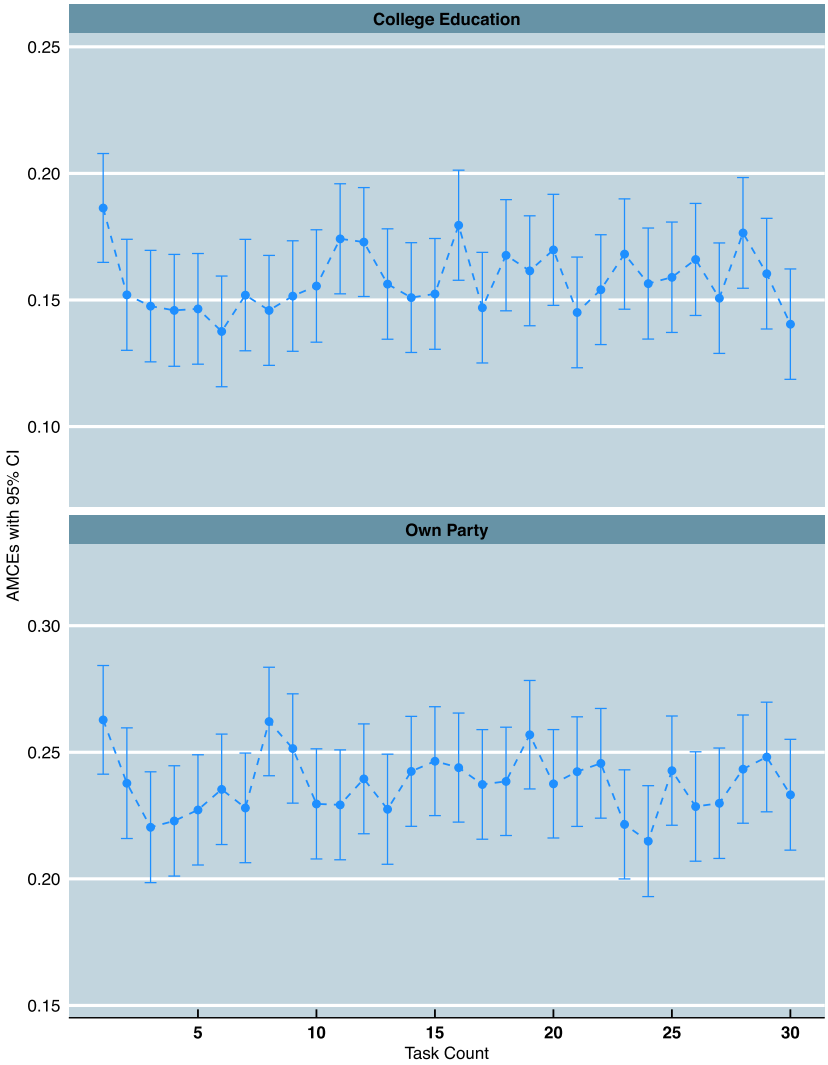

Figure 2. The AMCEs for our core attributes of interest from the five MT surveys as the number of completed choice tasks increases.

a large number of choice tasks. For both attributes, the estimated AMCEs are substantively large and statistically significant in the respondents' first task (0.186 and 0.263 for education and party, respectively, with $SE = 0.01$ for both attributes). The AMCEs then drop slightly for the second task (0.152 and 0.238, $SE = 0.01$ for both) but remain stable and close to that level throughout the duration of the survey, even occasionally jumping back to the original level. Even at the 30th task, the estimated AMCEs barely differ from those for the second task (0.140 and 0.233, $SE = 0.01$ for both). We note that the rate of sample attrition over the course of the 30 tasks is negligibly small, as is typical in surveys fielded on MT.

The result for the partial R^2 values, presented in Figure 3, confirms the stability of conjoint responses across the 30 tasks for our MT respondents. The partial R^2 for the two core attributes is about 0.104 in the respondents' first task, with a 95% block-bootstrapped confidence interval of [0.091, 0.118].[7] The coefficient drops slightly to 0.079 in the second task (with the 95% CI of [0.068, 0.092]) and remains remarkably stable around that value throughout the remaining 28 tasks. The two core attributes meaningfully explain the choice responses even at the very end of the lengthy

7 The block bootstrap procedure is based on resampling (with replacement) respondents rather than individual observations in order to account for the within-respondent correlation of the outcome variables.

https://doi.org/10.1017/pan.2017.40 Published online by Cambridge University Press

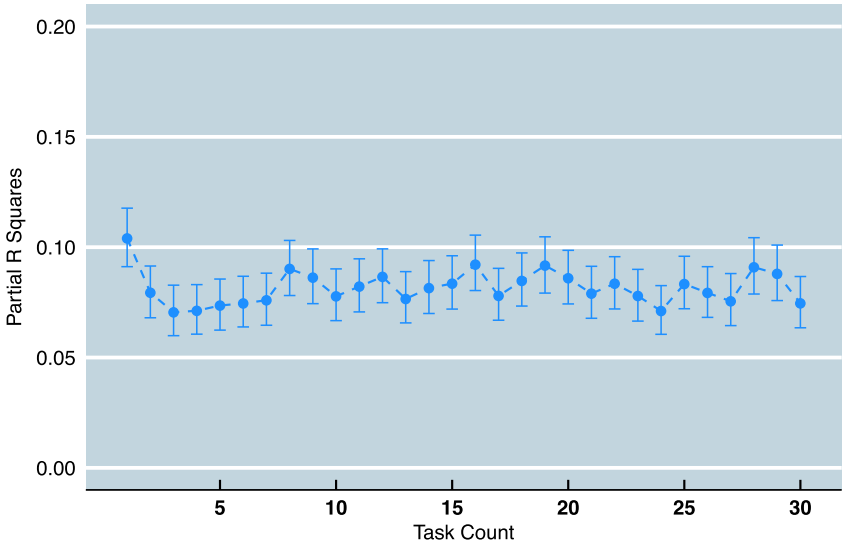

Figure 3. The partial R^2 values for our core attributes for the pooled MT data as a function of the number of completed tasks.

conjoint exercises ($R^2 = 0.075$, with 95% CI [0.063, 0.087]). These findings are replicated in our SSI sample, as shown in Figures A.1 and A.2 in Section A.2 of the supplementary materials.

In additional analyses reported in the supplementary materials, we also find that our results hold when evaluating the effects of other attributes included in our conjoint design, such as the candidates' age, military service, and policy positions (see Supplementary Figures A.3–A.6 in Section A.3). Overall, our study suggests that conjoint designs are remarkably impervious to threats from survey fatigue and satisfying when applied to respondents on MT and SSI, two of the most frequently used populations in experimental research.

4 Conclusion

The rapid growth of survey research conducted via computers has enabled researchers to employ increasingly complex research designs at little added cost. Conjoint experiments are one such design, and they have seen a renaissance within political science in the past few years. However, research on survey methods has to date been focused on the change in sampling frames that accompanies the shift toward online survey administration (e.g. Chang and Krosnick 2009; Yeager *et al.* 2011). For those administering surveys via computer, there is surprisingly little guidance about the extent to which insights developed for phone and in-person surveys hold up (but see Gooch and Vavreck 2015).

In this paper, we sought to advance our understanding of response behavior in surveys administered by computer by probing one breaking point of conjoint designs. Specifically, we considered an important decision confronting researchers who seek to implement conjoint experiments: how many tasks can one assign per respondent without inducing survey fatigue and excessive satisficing? Through a series of experiments, we find conjoint designs to be surprisingly robust, at least with the opt-in samples employed here (and in many other contemporary survey experiments). Even after completing 30 tasks, respondents continue to process the conjoint profiles in similar ways and to provide similar, sensible results.

These results allow us to make design recommendations for researchers interested in using conjoint survey experiments. While the results do not point to an optimal number of tasks, they show that the number of tasks is not a binding constraint for the experimental design in terms of satisficing—at least within the 30-task limit explored in this study. While we would not necessarily

https://doi.org/10.1017/pan.2017.40 Published online by Cambridge University Press

recommend researchers to use as many as 30 tasks, we have shown that within that limit, satisficing is not a serious concern that should dictate the number of tasks. Instead, researchers are free to make their decisions on the number of tasks on the basis of other design considerations, such as the survey length, cost constraints, and statistical power.

Certainly, the results from this study may differ for populations with little to no experience taking surveys via computer, or with reduced incentives to pay attention. The results may also differ in cases where the conjoint survey covers different subject matter. Making comparisons that are more familiar to respondents—such as between presidential candidates, job applicants, or consumer products—is likely to be easier than evaluating, for example, the elements of a complex policy proposal. Survey fatigue may be more pronounced and/or set in more quickly in a more complex context.

Conjoint experiments undoubtedly have breaking points—but our analyses suggest that at least for surveys administered with experienced and motivated survey takers, the breaking point in terms of the number of tasks appears to be beyond the range of common practice. Important questions remain about other aspects of conjoint design and their implications for survey response quality. In a companion study, we investigate the extent to which increasing the number of attributes in a conjoint design affects response quality (Bansak *et al.* 2017a). Broader questions include whether conjoint experiments might have advantages over alternative designs such as vignettes in terms of satisficing, to which Hainmueller, Hangartner, and Yamamoto (2015) provide some partial answers.

Supplementary material

For supplementary material accompanying this paper, please visit
https://doi.org/10.1017/pan.2017.40.

References

Abrajano, M. A., C. S. Elmendorf, and K. M. Quinn. 2015. *Using experiments to estimate racially polarized voting*. UC Davis Legal Studies Research Paper Series, no. 419. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2569982.

Acharya, A., M. Blackwell, and M. Sen. 2016. Analyzing causal mechanisms in survey experiments. July 5 Draft, Stanford University.

Bansak, K., J. Hainmueller, and D. Hangartner. 2016. How economic, humanitarian, and religious concerns shape European attitudes toward asylum seekers. *Science* 354(6309):217–222.

Bansak, K., J. Hainmueller, D. J. Hopkins, and T. Yamamoto. 2017a. *Beyond the breaking point? Survey satisficing in conjoint experiments*. Stanford University Graduate School of Business. Research Paper No 17-33; MIT Political Science Department Research Paper No. 2017-16.

Bansak, K., J. Hainmueller, D. J. Hopkins, and T. Yamamoto. 2017b. Replication data for: the number of choice tasks and survey satisficing in conjoint experiments. Harvard Dataverse, V1, UNF:6:DuDnb59vl0m5kHUWRTnYjg==, doi:10.7910/DVN/TLPMVI.

Bechtel, M., F. Genovese, and K. Scheve. 2017. Interests, norms, and support for the provision of global public goods: the case of climate co-operation. *British Journal of Political Science*, 1–23. doi:10.1017/S0007123417000205.

Berinsky, A. J., M. F. Margolis, and M. W. Sances. 2014. Separating the shirkers from the workers? making sure respondents pay attention on self-administered surveys. *American Journal of Political Science* 58(3):739–753.

Carnes, N., and N. Lupu. 2015. Do voters dislike politicians from the working class? Working Paper, Duke University.

Chang, L., and J. A. Krosnick. 2009. National surveys via rdd telephone interviewing versus the internet: comparing sample representativeness and response quality. *Public Opinion Quarterly* 73(4):641–678.

De Vaus, D. 2014. *Surveys in social research*. 6th Edition Routledge.

Egami, N., and K. Imai. 2015. Causal interaction in high dimension. Working paper, Princeton University.

Franchino, F., and F. Zucchini. 2015. Voting in a multi-dimensional space: a conjoint analysis employing valence and ideology attributes of candidates. *Political Science Research and Methods* 3(2):221–241.

Galesic, M., and M. Bosnjak. 2009. Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly* 73(2):349–360.

https://doi.org/10.1017/pan.2017.40 Published online by Cambridge University Press

Gooch, A., and L. Vavreck. 2015. How face-to-face interviews and cognitive skill affect non-response: a randomized experiment assigning mode of interview. Working Paper, University of California, Los Angeles.
Green, P. E., and V. R. Rao. 1971. Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research* VIII:355–363.
Groves, R. M., F. J. Fowler Jr., M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. 2011. *Survey methodology*, vol. 561, John Wiley & Sons.
Hainmueller, J., D. Hangartner, and T. Yamamoto. 2015. Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences* 112(8):2395–2400.
Hainmueller, J., and D. J. Hopkins. 2015. The hidden American immigration consensus: a conjoint analysis of attitudes toward immigrants. *American Journal of Political Science* 59(3):529–548.
Hainmueller, J., D. J. Hopkins, and T. Yamamoto. 2014. Causal inference in conjoint analysis: understanding multidimensional choices via stated preference experiments. *Political Analysis* 22(1):1–30.
Hainmueller, J., D. J. Hopkins, and T. Yamamoto. 2015. Learning more from conjoint experiments through a doubly randomized design. Paper presented at the Annual Meeting of APSA.
Herzog, A. R., and J. G. Bachman. 1981. Effects of questionnaire length on response quality. *Public Opinion Quarterly* 45(4):549–559.
Horiuchi, Y., D. M. Smith, and T. Yamamoto. 2015. Identifying multidimensional policy preferences of voters in representative democracies: a conjoint field experiment in Japan. Working Paper, MIT.
Jasso, G., and P. H. Rossi. 1977. Distributive justice and earned income. *American Sociological Review* 42(4):639–651.
Kahneman, D. 2003. A perspective on judgment and choice: mapping bounded rationality. *American Psychologist* 58(9):697–720.
Krosnick, J. A. 1999. Survey research. *Annual Review of Psychology* 50(1):537–567.
Mummolo, J., and C. Nall. 2017. Why partisans don't sort: the constraints on political segregation. *The Journal of Politics* 79(1):45–59.
Raghavarao, D., J. B. Wiley, and P. Chitturi. 2011. *Choice-based conjoint analysis: models and designs*. Boca Raton, FL: CRC Press.
Stewart, N., C. Ungemach, A. J. Harris, D. M. Bartels, B. R. Newell, G. Paolacci, and J. Chandler. 2015. The average laboratory samples a population of 7,300 amazon mechanical turk workers. *Judgment and Decision Making* 10(5):479–491.
Wright, M., M. Levy, and J. Citrin. 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.
Yeager, D. S., J. A. Krosnick, L. Chang, H. S. Javitz, M. S. Levendusky, A. Simpser, and R. Wang. 2011. Comparing the accuracy of rdd telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly* 75(4):709–747.

https://doi.org/10.1017/pan.2017.40 Published online by Cambridge University Press