CHAPTER 2

Conjoint Survey Experiments*

Kirk Bansak, Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto

Abstract

Conjoint survey experiments have become a popular method for analyzing multidimensional preferences in political science. If properly implemented, conjoint experiments can obtain reliable measures of multidimensional preferences and estimate causal effects of multiple attributes on hypothetical choices or evaluations. This chapter provides an accessible overview of the methodology for designing, implementing, and analyzing conjoint survey experiments. Specifically, we begin by detailing a substantive example: How do candidate attributes affect the support of American respondents for candidates running against President Trump in 2020? We then discuss the theoretical underpinnings and advantages of conjoint designs. We next provide guidelines for practitioners in designing and

analyzing conjoint survey experiments. We conclude by discussing further design considerations, common conjoint applications, common criticisms, and possible future directions.

2.1 Introduction

Political and social scientists are frequently interested in how people choose between options that vary in multiple ways. For example, a voter who prefers candidates to be experienced and opposed to immigration may face a dilemma if an election pits a highly experienced immigration supporter against a less experienced immigration opponent. One might ask similar questions about a wide range of substantive domains – for instance, how people choose whether and whom to date, which job to take, and where to rent or buy a home. In all of these examples, and in many more, people must choose among multiple options that are themselves collections of attributes. In making such choices, people must not only identify their preferences on each particular dimension, but also make trade-offs across the dimensions.

Conjoint analysis is a survey-experimental technique that is widely used as a tool to answer these types of questions across the social sciences. The term originates in the study of "conjoint measurement" in 1960s mathematical psychology, when founding figures in the behavioral sciences such as Luce and Tukey (1964) developed axiomatic theories for decomposing "complex phenomena into sets of basic factors according to specifiable rules of combination" (Tversky 1967). Since the seminal publication of Green and Rao (1971), however, the term "conjoint analysis" has primarily been

used to refer to a class of survey-experimental methods that estimates respondents' preferences given their overall evaluations of alternative profiles that vary across multiple attributes, typically presented in tabular form.

Traditional conjoint methods drew heavily on the statistical literature on the design of experiments (DOE) (e.g., Cox 1958), in which theories of complex factorial designs were developed for industrial and agricultural applications. However, conjoint designs became especially popular in marketing (see Raghavarao et al. 2011), as it was far easier to have prospective customers evaluate hypothetical products on paper than to build various prototypes of cars or hotels. Conjoint designs were also frequently employed in economics (Adamowicz et al. 1998) and sociology (Jasso and Rossi 1977; Wallander 2009), often under different names such as "stated choice methods" or "factorial surveys." In the era before computer-assisted survey administration, respondents would commonly have to evaluate dozens of hypothetical profiles printed on paper, and even then, analysis proceeded under strict assumptions about the permissible interactions among the attributes.

Only in recent years, however, have conjoint survey experiments come to see extensive use in political science (e.g., Abrajano et al. 2015; Bansak et al. 2016; Bechtel et al. 2019; Carnes and Lupu 2016; Franchino and Zucchini 2015; Hainmueller and Hopkins 2015; Horiuchi et al. 2018; Lowen et al. 2012; Mummolo and Nall 2016; Wright et al. 2016). This development has been driven partly by the proliferation of computer-administered surveys and by the concurrent ability to conduct fully randomized conjoint experiments at low cost. Reflecting the explosion of conjoint applications in academic

political science publications, a conjoint analysis of Democratic voters' preferences for presidential candidates even made an appearance on television via CBS News in the spring of 2019 (Khanna 2019). A distinctive feature of this strand of empirical literature is a new statistical approach to conjoint data based on the potential outcomes framework of causal inference (Hainmueller et al. 2014), which is in line with the explosion in experimental methods in political science generally since the early 2000s (Druckman et al. 2011). Along with this development, the past several years have also seen valuable advances in the statistical methods for analyzing conjoint data that similarly build on modern causal inference frameworks (Acharya et al. 2018; Dafoe et al. 2018; Egami and Imai 2019).

In this chapter, we introduce conjoint survey experiments, summarize recent research employing them and improving their use, and discuss key issues that emerge when putting them to use. We do so partly through the presentation and discussion of an original conjoint application in which we examine an opt-in sample of Americans' attitudes toward prospective 2020 Democratic presidential nominees.

2.2   An Empirical Example: Candidates Running against President Trump in 2020

To illustrate how one might implement and analyze a conjoint survey experiment, we conducted an original survey on an online, opt-in sample of 503 Amazon Mechanical Turk (MTurk) workers (for discussion of these types of samples, see the Krupnikov, Nam, and Style chapter in this volume). We designed our experiment to be illustrative of a typical conjoint design in political

science. Specifically, we presented respondents with a series of tables showing profiles of hypothetical Democratic candidates running in the 2020 US presidential election. We asked: "This study is about voting and about your views on potential Democratic candidates for President in the upcoming 2020 general election ... please indicate which of the candidates you would *prefer to win the Democratic primary* and hence *run against President Trump* in the general election" (emphasis in the original). We then presented a table that contained information about two political candidates side by side, described as "CANDIDATE A" and "CANDIDATE B," which were purported to represent hypothetical Democratic candidates for the 2020 election. Figure 2.1 shows an example table from the experiment.

	CANDIDATE A	CANDIDATE B
Military Service Experience	Served in the Navy	Did not serve
Supports Creating Pathway to Citizenship for	Unauthorized immigrants with no criminal record who entered the U.S. as minors	All unauthorized immigrants with no criminal record
Previous Occupation	Lawyer	Lawyer
Age	53	45
Gender	Female	Female
Race/Ethnicity	White	Hispanic/Latino
Sexual Orientation	Gay	Straight
Position on Climate Change	Ban the use of fossil fuels after 2040, substantially reducing economic growth	Impose a tax on using fossil fuels, moderately reducing economic growth
Supports Government Healthcare for	Only Americans who are older, poor, or disabled	Only Americans who are older, poor, or disabled
Prior Political Experience	U.S. Senator	No prior political experience

Figure 2.1 An example conjoint table from the Democratic primary experiment. The full set of possible attribute values is provided in Table 2.1.

As is shown in Figure 2.1, conjoint survey experiments typically employ a tabular presentation of multiple pieces of information representing various attributes of hypothetical objects. This table is typically referred to as a "conjoint table" since it combines a multitude of varying attributes and presents them as a single object. In our experiment, we used a table containing two profiles of hypothetical Democratic candidates varying in terms of their age, gender, sexual orientation, race/ethnicity, previous occupation, military service, prior political experience, and positions on healthcare policy, immigration policy, and climate change policy. Table 2.1 shows the full set of possible levels for each of the attributes. We worked to choose a range of attributes that would be likely to be salient to voters during actual choices among primary candidates, but of course the conjoint presentation will make select attributes more salient than their real-world counterparts while ignoring others.

Table 2.1   The list of possible attribute values in the Democratic primary experiment.

Age	37, 45, 53, 61, 77
Gender	Female, Male
Sexual Orientation	Straight, Gay
Race/Ethnicity	White, Hispanic/Latino, Black, Asian
Previous Occupation	Business executive, College professor, High school teacher, Lawyer, Doctor, Activist

"conjoint table" since it combines a multitude of varying attributes and presents them as a single object. In our experiment, we used a table containing two profiles of hypothetical Democratic candidates varying in terms of their age, gender, sexual orientation, race/ethnicity, previous occupation, military service, prior political experience, and positions on healthcare policy, immigration policy, and climate change policy. Table 2.1 shows the full set of possible levels for each of the attributes. We worked to choose a range of attributes that would be likely to be salient to voters during actual choices among primary candidates, but of course the conjoint presentation will make select attributes more salient than their real-world counterparts while ignoring others.

Table 2.1 The list of possible attribute values in the Democratic primary experiment.

Age	37, 45, 53, 61, 77
Gender	Female, Male
Sexual Orientation	Straight, Gay
Race/Ethnicity	White, Hispanic/Latino, Black, Asian
Previous Occupation	Business executive, College professor, High school teacher, Lawyer, Doctor, Activist
Military Service Experience	Did not serve, Served in the Army, Served in the Navy, Served in the Marine Corps
Prior Political Experience	Small-city Mayor, Big-city Mayor, State Legislator, Governor, U.S. Senator, U.S. Representative, No prior political experience
Supports Government Healthcare for	All Americans, Only Americans who are older, poor, or disabled, Americans who choose it over private health plans

Supports Creating Pathway to Citizenship for	Unauthorized immigrants with no criminal record who entered the U.S. as minors, All unauthorized immigrants with no criminal record, No unauthorized immigrants
Position on Climate Change	Ban the use of fossil fuels after 2040, reducing economic growth by 5%; Impose a tax on using fossil fuels, reducing economic growth by 3%; Promote the use of renewable energy but allow continued use of fossil fuels

The levels presented in each table were then randomly varied, with randomization occurring independently across respondents, across tables, and across attributes. Each respondent was presented 15 such randomly generated comparison tables on separate screens, meaning that they evaluated a total of 30 hypothetical candidates (i.e., 15 choices between candidates).

In order to preserve a smooth survey-taking experience, the order in which attributes were presented was held fixed across all 15 tables for each individual respondent, though the order was randomized across respondents. Put differently, every respondent saw the attributes in the same order for each of the 15 scenarios, but that order randomly varied across respondents.

After presenting each of the conjoint tables with randomized attributes, we asked respondents two questions to measure their preferences about the hypothetical candidate profiles just presented. Specifically, we used a seven-point rating of the profiles (top of Figure 2.2) and a forced choice between the two profiles (bottom of Figure 2.2). We asked: "On a scale from 1 to 7 … how would you rate each of the candidates described above?" and also: "Which candidate profile would you prefer for the Democratic candidate to run against President Trump in the general election?" The order of these two items was randomized (at the respondent level) so that we would be able to identify any order effects on outcome measurement if necessary.

Figure 2.2 Outcome variables in the Democratic primary experiment.

The substantive goal of our conjoint survey experiment was twofold and can be encapsulated by the following questions. First, what attributes causally increase or decrease the appeal of a Democratic primary candidate, on average, when varied independently of the other candidate attributes included in the design? As we discuss later in the chapter, the random assignment of attribute levels allows researchers to answer this question by estimating a causal effect called the average marginal component effect (AMCE) using simple statistical methods such as linear regression. Second, do the effects of the attribute vary depending on whether the respondent is a Democrat, Republican, or independent? For respondents who are Democrats, the conjoint task simulated the choice of their own presidential candidate to run against President Trump in the 2020 presidential election. So the main trade-off for them was whether to choose a candidate who was electable or a candidate who represented their own policy positions more genuinely. On the other hand, for Republican respondents, considerations were likely to be entirely different (at least for those who intended to vote for President Trump). As we show later, these questions can be answered by estimating conditional AMCEs (i.e., the average effects of the attributes conditional on a respondent characteristic measured in the survey, such as partisanship).

2.3 Advantages of Conjoint Designs over Traditional Survey Experiments

Our Democratic primary experiment represents a typical example of the conjoint survey

experiments widely implemented across the empirical subfields of political science. A few factors have driven the upsurge in the use of conjoint survey experiments. First, there has been increased attention to causal inference and to experimental designs that allow for inferences about causal effects via assumptions made credible by the experimental design itself (Sniderman and Grob 1996). At the same time, however, researchers are often interested in testing hypotheses that go beyond the simple cause-and-effect relationship between a single binary treatment and an outcome variable. Traditional survey experiments are typically limited to analyzing the average effects of a few randomly assigned treatments, constraining the range of substantive questions researchers can answer persuasively. In contrast, conjoint experiments allow researchers to estimate the effects of many attributes simultaneously, and so can permit analysis of more complex causal questions.

A second enabling factor is the rapid expansion of surveys administered via computer, which enables researchers to use fully randomized conjoint designs (Hainmueller et al. 2014). Fully randomized designs, in turn, facilitate the estimation of key quantities such as AMCEs via straightforward statistical estimation procedures that rely little on modeling assumptions. Moreover, commonly used web-based survey interfaces facilitate the implementation of complex survey designs such as conjoint experiments.

A third underlying factor behind the rise of conjoint designs within political science is their close substantive fit with key political science questions. For example, political scientists have long been interested in how voters choose among candidates or parties, a question for which conjoint designs are well suited. By quantifying the causal effects of various candidate attributes presented simultaneously, conjoint designs enable researchers to explore a wide range of hypotheses about voters' preferences, relative sensitivities to different attributes, and biases. But beyond voting, multidimensional choices and preferences are of interest to political scientists in many contexts and issue areas, such as immigration, neighborhoods and housing, and regulatory policy packages. As we discuss later in this chapter, conjoint designs have been applied

in each of these domains and beyond.

Fourth, political scientists are often interested in measuring attitudes and preferences that might be subject to social desirability bias. Scholars have argued that conjoint designs can be used as effective measurement tools for socially sensitive attitudes, such as biases against female political candidates (Teele et al. 2018) and opposition to siting a low-income housing project in one's neighborhood (Hankinson 2018). When respondents evaluate several attributes simultaneously, they may be less concerned that researchers will connect their choices to one specific attribute. In keeping with this expectation, early evidence suggests that fully randomized conjoint designs do indeed mitigate social desirability bias by asking about a socially sensitive attribute along with a host of other randomly varying attributes (Horiuchi et al. 2019).

Finally, evidence suggests that conjoint designs have desirable properties in terms of validity. On the dimension of external validity, Hainmueller et al. (2015) find that certain conjoint designs can effectively approximate real-world benchmarks in Swiss citizenship votes, while Auerbach and Thachil (2018) find that political brokers in Indian slums have the attributes that local residents reported valuing via a conjoint experiment. Conjoint designs have also proven to be quite robust. For one thing, online, opt-in respondents commonly employed in social science research can complete many conjoint tasks before satisficing demonstrably degrades response quality (Bansak et al. 2018). Such respondents also prove able to provide meaningful and consistent responses even in the presence of a large number of attributes (Bansak et al. 2019; see also Jenke et al. 2020).

In short, conjoint designs have a range of theoretical and applied properties that make them attractive to political scientists. But, of course, no method is appropriate for all applications. Later in this chapter, we therefore flag the limitations of conjoint designs as well as the open questions about their usage and implementation.

2.4 Designing Conjoint Survey Experiments

When implementing a conjoint experiment,

survey experimentalists who are new to conjoint analysis face a multitude of design considerations. Here, we review a number of key components of a conjoint design that have implications for conjoint measurement and offer guidance on how to approach them, using the Democratic primary experiment as a running example.

2.4.1 Number of Profiles

In the Democratic primary experiment, we used a "paired-profile" design in which each conjoint table contained two profiles of hypothetical Democratic candidates. But other designs are also possible. One example is a "single-profile" design in which each table presents only one set of attribute values; another is a multiple-profile design that contains more than two profiles per table. Empirically, paired-profile designs appear to be the most popular choice among political scientists, followed by single-profile designs. Hainmueller et al. (2015) provide empirical justification for this choice, showing that paired-profile designs tend to perform well compared to single-profile designs, at least in the context of their study of Swiss voting on naturalization. In other contexts, a single profile or more than two profiles may be most appropriate, subject of course to limitations in respondents' ability to compare many profiles simultaneously.

2.4.2 Number of Attributes

An important practical question is how many attributes to include in a conjoint experiment. Here, researchers face a difficult trade-off between *masking* and *satisficing* (Bansak et al. 2019). On the one hand, including too few attributes will make it difficult to interpret the substantive meaning of AMCEs, since respondents might associate an attribute with another that is omitted from the design. Such a perceived association between an attribute included in the design and another omitted attribute muddies the interpretation of the AMCE of the former as it may represent the effects of both attributes (i.e., masking; for more, see Bansak et al. 2019; Dafoe et al. 2018). In our Democratic primary experiment, for example, the AMCEs of the policy position attributes might mask the effect of other policy positions that are not included in the design if respondents associate a liberal position on the former with a

similarly liberal position on the latter. On the other hand, including too many attributes might increase the cognitive burden of the tasks excessively, inducing respondents to satisfice (Krosnick 1999).

Given the inherent trade-off, how many attributes should one use in a conjoint experiment? Although the answer to the question is likely to be highly context dependent, Bansak et al. (2019) provide useful evidence that subjects recruited from popular online survey platforms such as MTurk are reasonably resistant to satisficing due to the increase in the number of conjoint attributes. Based on the evidence, they conclude that the upper bound on the permissible number of conjoint attributes for online surveys is likely to be above those used in typical conjoint experiments in political science, such as our Democratic primary example in which 10 attributes were used. Jenke et al. (2020) also explore the robustness of conjoint experiments to the addition of attributes by using eye-tracking methods to examine how respondents process information in conjoint surveys, administered to university students and local community members. They find that respondents are able to adapt to the increased complexity of additional attributes and to reduce cognitive processing costs by selectively incorporating relevant new information into their choices (and ignoring less relevant information). Of course, how many attributes might be too many also likely depends on the sample of respondents and the mode of delivery.

2.4.3 Randomization of Attribute Levels

Regardless of the number of profiles per table, conjoint designs entail a random assignment of attribute values. The canonical, fully randomized conjoint experiment randomly draws a value for each attribute in each table from a prespecified set of possible values (Hainmueller et al. 2014). This makes the fully randomized conjoint experiment a particular type of factorial experiment, on which an extensive literature exists in the field of DOE. In our experiment, for example, we chose the set of possible values for the age attribute to be [33, 45, 53, 61, 77], and we randomly picked one of these values for each profile with equal probability (= 1/5). As discussed later, the random assignment of

attribute values enables inference about the causal effects of the attributes without reliance on untestable assumptions about the form of respondents' utility functions or the absence of interaction effects (Hainmueller et al. 2014).[1]

In most existing applications of conjoint designs in political science, attributes are randomized uniformly (i.e., with equal probabilities for all levels in a given attribute) and independently from one another. Although uniform independent designs are attractive because of parsimony and ease of implementation, the conjoint design can accommodate other kinds of randomization distributions. Often, researchers have good reasons to deviate from the standard uniform independent design for the sake of realism and external validity (Hainmueller et al. 2014). In designing our experiment, for example, we wanted to ensure that the marginal distributions of the candidate attributes were roughly representative of the attributes of the politicians who were considered to be likely candidates in the actual Democratic primary election at that time. Thus, in addition to choosing attribute values that matched those of the actual likely candidates, we employed a weighted randomization such that some values would be drawn more frequently than others. Specifically, we made our hypothetical candidates more likely: to be straight than gay (with 4:1 odds); to be White than Black, Latino/Hispanic, or Asian (6:2:2:1); and to have never served in the military than to have served in the Army, Navy, or Marine Corps (4:1:1:1). Weighted randomization causes no fundamental threat to the validity of causal inference in conjoint analysis, although it introduces some important nuances in the estimation and interpretation of the results. We will come back to these issues in the next section.

Another possible "tweak" to the randomization distribution is to introduce dependence between some attributes (Hainmueller et al. 2014). The most common instance of this is restricted randomization, or prohibiting certain combinations of attribute values from happening. Restricted randomization is typically employed to ensure that respondents will not encounter completely unrealistic (or sometimes even logically impossible) profiles. For example, in the "immigration conjoint" study reported in Hainmueller et al. (2014), the authors impose the

restriction that immigrants with high-skilled occupations must at least have a college degree. In our current Democratic primary experiment, we chose not to impose any such "hard" constraints on the randomization distribution because we chose attribute values that were all reasonably plausible to co-occur in an actual profile of a Democratic candidate. Like weighted randomization, restricted randomization does not pose a fundamental problem for making valid causal inferences from conjoint experiments, unless it is taken to the extreme. However, restricted designs require care in terms of estimation and interpretation, especially when it is not clear what combinations of attributes make a profile unacceptably unrealistic. More discussion is found later in this chapter.

2.4.4   Randomization of Attribute Ordering

In addition to randomizing the values of attributes, it is often recommended to randomize the order of the attributes in a conjoint table, so that the causal effects of attributes themselves can be separately identified from pure order effects (e.g., the effects of an attribute being placed near the top of the table vs. towards the bottom). In many applications, attribute ordering is better randomized at the respondent level (i.e., for a given respondent, randomly order attributes in the first table and fix the order throughout the rest of the experiment), and that is precisely what we did in the experiment presented here. This is because reshuffling the order of attributes from one table to another is likely to cause excessive cognitive burden for respondents (Hainmueller et al. 2014).

2.4.5   Outcome Measures

After presenting a conjoint table with randomized attributes, researchers then typically ask respondents to express their preferences with respect to the profiles presented. These preferences can be measured in various ways, and those measurements then constitute the outcome variable of interest in the analysis of conjoint survey data. The individual rating and forced choice outcomes are the two most common measures of stated preference in political science applications of conjoint designs, and there are distinct advantages to each. On the one hand, presenting a forced choice may compel

respondents to think more carefully about trade-offs. On the other hand, individual ratings (or non-forced choices where respondents can accept/reject all profiles presented) allow respondents to express approval or disapproval of each profile without constraints, which also allows for the identification of respondents that categorically accept/reject all profiles.[2]

It is important to note that whether respondents are forced to choose among conjoint profiles or are able to rate them individually can influence one's conclusions, so it is often valuable to elicit preferences about profiles in multiple ways. Indeed, researchers commonly ask respondents to both rank profiles within a group and to rate each profile individually.

2.4.6 Number of Tasks

In typical conjoint survey experiments in political science, the task (i.e., a randomly generated table of profiles followed by outcome measurements) is repeated multiple times for each respondent, each time drawing a new set of attribute values from the same randomization distribution. In our Democratic primary experiment, respondents were given 15 paired comparison tasks, which

means they evaluated a total of 30 hypothetical candidate profiles. One important advantage of conjoint designs is that one can obtain many more observations from a given number of respondents without compromising validity than a traditional survey experiment, where within-subject designs are often infeasible due to validity concerns. This, together with the fact that one can also test the effects of a large number of attributes (or, equivalently, treatments) at once, makes the conjoint design a highly cost-efficient empirical strategy. One concern, however, is the possibility of respondent fatigue when the number of tasks exceeds respondents' cognitive capacity.

The question then is: How many tasks are too many? The answer is likely highly dependent on the nature of the conjoint task (e.g., how complicated the profiles are) and of the respondents (e.g., how familiar they are with the subject matter at hand), so it is wise to err on the conservative side. However, Bansak et al. (2018) empirically show that inferences from conjoint designs are robust with respect to the number of tasks for samples recruited from commonly used online opt-in panels. In particular, their findings

indicate that it is safe to use as many as 30 tasks with respondents from MTurk and Survey Sampling International's online panel without detectable degradation in response quality. Although one should be highly cautious in extrapolating their findings to other samples, it appears to reinforce the use of 15 tasks in our Democratic primary experiment, which draws on MTurk respondents. To be sure, researchers need to consider the overall survey length and its effects on attrition and respondent engagement as well.

2.4.7 Variants of Conjoint Designs

Finally, a survey experimental design that is closely related to the conjoint experiment is the so-called vignette experiment. Like a conjoint experiment, a vignette experiment typically describes a hypothetical object that varies in terms of multiple attributes and asks respondents to either rate or choose their preferred profiles. The key difference is that a profile is presented as a descriptive text as opposed to a table. For example, a vignette version of our Democratic primary experiment would use a paragraph like the following to describe the profile of a candidate: "CANDIDATE A is a 37-year-old straight Black man with no past military service or political experience. He used to be a college professor. He supports providing government healthcare for all Americans, creating a pathway to citizenship for unauthorized immigrants with no criminal record, and a complete fossil fuel ban after 2040 even with a substantial reduction in economic growth."

The vignette design can simply be viewed as a type of a conjoint experiment, since it shares most of the key design elements with table-based conjoint experiments that we have assumed in our discussion so far. However, there are a few important reasons to prefer the tabular presentation of attributes in many cases. First, it can be more difficult to randomize the order of attributes in a vignette experiment, since certain changes might cause the text to become incoherent due to grammatical and sentence structure issues. Second, Hainmueller et al. (2015) show empirically that, at least in their validation study, vignette designs tend to perform less well than tabular conjoint designs, and they also find evidence suggesting that the performance advantage for tabular conjoint

designs is due to increased engagement with the survey. Specifically, they find that the effects estimated from a vignette design are consistently attenuated towards zero (while maintaining the directions) compared to the estimates from an otherwise identical tabular conjoint experiment. Vignettes may also heighten respondent fatigue and so reduce the number of tasks respondents are able to complete without excessive satisficing. That being said, certain research questions might naturally call for a vignette format, and the analytical framework discussed below is directly applicable to fully randomized vignette designs as well.

2.5 Analyzing Data from Conjoint Survey Experiments

In this section, we provide an overview of the common statistical framework for the causal analysis of conjoint survey data. Much of the theoretical underpinning for the methodology comes directly from the literature on potential outcomes and randomized experiments (e.g., Imbens and Rubin 2015). We refer readers to Hainmueller et al. (2014) for a more formal

treatment of the materials here.

A key quantity in the analysis of conjoint experiments is the AMCE, a causal estimand first defined by Hainmueller et al. (2014) as a quantity of interest. Our discussion below thus focuses on what the AMCE is, how it can be estimated, and how to interpret it. Given the interest among political scientists in using conjoint experiments to study elections, it is worth highlighting that the AMCE, when applied to elections, is directly interpretable as a causal effect on a candidate's or party's expected vote share. For a more detailed discussion of the AMCE in this context, see Bansak et al. (2020).

2.5.1 Motivation, Definition, and Estimation

As we discussed in the previous section, the fully randomized conjoint design is a particular instance of a full factorial design, where each of the attributes can be thought of as a multi-valued factor (or a "treatment component" in our terminology). This enables us to analyze conjoint survey data as data arising from a survey experiment with multiple randomized categorical

treatments, to which we can apply a standard statistical framework for causal inference such as the potential outcomes framework.[3] From this perspective, the analysis of conjoint survey data is potentially straightforward, for the average treatment effect (ATE) of any particular combination of the treatment values against another can be unbiasedly estimated by simply calculating the difference in the means of the observed outcomes between the two groups of responses that were actually assigned those treatment values in the data. For example, in our Democratic primary experiment, we might consider estimating the ATE of a 61-year-old straight White female former business executive with no prior military service or experience in elected office who supports government-provided healthcare for all Americans, creating pathway to citizenship for all unauthorized immigrants with no criminal record, and imposing a tax on fossil fuels, versus a 37-year-old gay Latino male former lawyer turned state legislator with no military service who supports the same positions on healthcare, unauthorized immigrants, and climate change.

Thinking through this example immediately makes it apparent that this approach has several problems. First, substantively, researchers rarely have a theoretical hypothesis that concerns a contrast between particular pairs of attribute value combinations when their conjoint table includes as many attributes as in our experiment. Instead, researchers employing a conjoint design are typically primarily interested in estimating the effects of individual attributes, such as the effect of gender, while allowing respondents also to consider explicitly other attributes that might affect their evaluations of the hypothetical candidates. In other words, a typical quantity of interest in conjoint survey experiments is the overall effect of a given attribute averaged across other attributes that also appear in the conjoint table.

Second, statistically, estimating the effect of a particular combination of attribute values against another based on a simple difference in means requires an enormous sample size, since the number of possible combinations of attribute values is very large compared to the number of actual observations. In our experiment, there were $5 \times 2 \times 2 \times 3 \times 6 \times 4 \times 6 \times 3 \times 3 \times 3 = 233{,}280$ possible unique profiles, whereas our

observed data contained only 30 × 503 = 15,090 sampled profiles. This implies that observed data from a fully randomized conjoint experiment are usually far too sparse to estimate the ATEs of particular attribute combinations for the full set of attributes included in the study.

For these reasons, researchers instead focus on an alternative causal quantity called the AMCE in most applications of conjoint survey experiments in political science. The AMCE represents the effect of a particular attribute value of interest against another value of the same attribute while holding equal the joint distribution of the other attributes in the design, averaged over this distribution as well as the sampling distribution from the population. This means that an AMCE can be interpreted as a summary measure of the overall effect of an attribute after taking into account the possible effects of the other attributes by averaging over the effect variations caused by them. For example, suppose that one is interested in the overall effect on the rating outcome measure of a candidate being female as opposed to male in our Democratic primary experiment. That is, what is the average causal effect of being a female candidate as opposed to a male candidate on the respondent's candidate rating when they are also given information about the candidate's age, race/ethnicity, etc.? To answer this question, one can estimate the AMCE of female versus male by simply calculating the average rating of all realized female candidate profiles, calculating the average rating of all male profiles, and taking the difference between the two averages.[4] Put differently, the AMCE is an average over the distribution of other attributes. The same procedure could also be performed with respect to the forced choice outcome measure to assess the average causal effect of being a female candidate as opposed to a male candidate on the probability that a candidate will be chosen. In that case, one can estimate the AMCE of female versus male by calculating the proportion of all realized female candidate profiles that were chosen, calculating the proportion of all male profiles that were chosen, and taking the difference between the two. The fact that the AMCE summarizes the overall average effect of an attribute when respondents are also given information on other attributes is appealing substantively because in reality respondents would often have such information

on other attributes when making a multidimensional choice.

2.5.2 Interpretation

Figure 2.3 shows estimated AMCEs for each of the 10 attributes included in our Democratic primary experiment along with their 95% confidence intervals, using the forced choice item as the outcome measure. These AMCEs were estimated via linear regression. Interpreting AMCEs is intuitive. For example, for our opt-in sample of 503 American respondents recruited through MTurk, presenting a hypothetical candidate as straight as opposed to gay increased the probability of respondents choosing the profile as their preferred candidate by about 4 percentage points on average when respondents were also given information about the other nine attributes. Thus, the AMCE represents a causal effect of an attribute value against another averaged over possible interaction effects with the other included attributes, as well as over possible heterogenous effects across respondents.
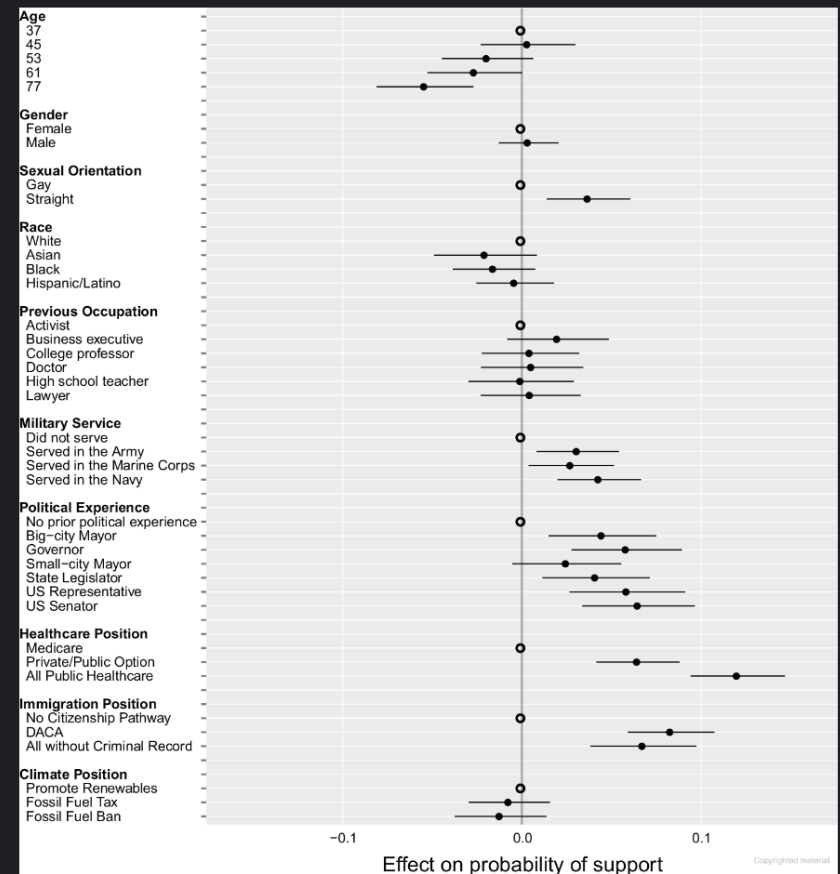

Figure 2.3 Average marginal component effects of candidate attributes in the Democratic primary conjoint experiment (forced choice outcome). DACA = Deferred Action for Childhood Arrivals.

Despite its simplicity, there are important nuances to keep in mind when interpreting AMCEs that are often neglected in applied research. First, the AMCE of an attribute value is always defined with respect to a particular

reference value of the same attribute, or the "baseline" value of the attribute. This is parallel to any other regression model or a standard survey experiment in which a treatment effect always represents the effect of the treatment against the particular control condition used in the experiment. Researchers sometimes neglect this feature when analyzing conjoint experiments, as Leeper et al. (2020) point out.

Second, an important feature of the AMCE as a causal parameter is that it is always defined with respect to the distribution used for the random assignment of the attributes. That is, the true value of the AMCE, as well as its substantive meaning, also changes when one changes the randomization distribution, unless the effect of the attribute has no interaction with other attributes. For example, as mentioned earlier, we used nonuniform randomization distributions for assigning some of the candidate attributes in our Democratic primary experiment, such as candidates' sexual orientation. Had we used a uniform randomization for the sexual orientation attribute (i.e., 1/2 straight and 1/2 gay) instead, the AMCE of another attribute (e.g., gender) could have been either larger or smaller than

what is reported in Figure 2.3, depending on how the effect of that attribute interacts with that of sexual orientation. This important nuance should always be kept in mind when interpreting AMCEs. Hainmueller et al. (2014) discuss this point in more detail (see also de la Cuesta et al. 2019).

As a related point, the AMCE differs in its handling of ties from certain other estimation strategies (such as conditional logistic regression) that have sometimes been applied to analyze forced choices (see also Ganter 2019). Consider an example in which a respondent is required to choose between two profiles with two attributes each, including one attribute of interest that has two levels, each appearing with 0.5 probability. Let's further assume that respondents care only about the attribute of interest, and they flip a coin to decide if the profiles are identical on that attribute. In 50% of all pairings, the levels of this attribute of interest will indeed be equal in the two profiles. As a consequence, even a respondent who always prefers one level of that attribute of interest will commonly see pairings in which both profiles have that attribute level, and so will be forced to

choose against profiles that include their preferred attribute level. Conversely, they will also see some profile pairings in which their preferred attribute level is entirely absent, and so will have to choose the non-preferred attribute level. The upshot is that because of the presence of tied attribute levels, the AMCEs of interest in this example will be −0.5 and 0.5.

More generally, in the presence of ties, the AMCEs for forced choice outcomes will be bounded at strictly less than 1 or greater than −1, with the extent of the deviation from −1 and 1 increasing with the probability of ties. The intuition is straightforward: even someone who always prefers female candidates will have to choose a male candidate in any pairing that pits two male profiles against each other. This issue does not emerge when respondents are rating individual profiles. Because ties occur frequently in many real-world settings, retaining ties will typically increase the realism of the conjoint. In some instances, researchers may wish to use constrained randomization to avoid ties altogether or to produce ties deliberately, such as when studying social desirability (e.g., Horiuchi et al. 2019). But when doing so, it is critical to acknowledge that such constrained randomizations change the distribution of attribute levels over which the other AMCEs are defined.

Finally, it is worth reiterating that the AMCE represents an *average* of individual-level causal effects of an attribute. In other words, for some respondents the attribute might have a large effect and for others the effect might be small or zero, and the AMCE represents the average of these potentially heterogeneous effects. This is no different from most of the commonly used causal estimands in any other experiment, such as the ATE or local ATE. Researchers often care about average causal effects because they provide an important and concise summary of what would happen on average to the outcome if everybody moved from control to treatment (Holland 1986). The fact that the ATE and AMCE average over both the sign and the magnitude of the individual-level causal effects is an important feature of these estimands, because both sign and magnitude are important in capturing the response to a treatment. As a case in point, one of the only real-world empirical validations of conjoint experiments of which we are aware finds

evidence that AMCEs from a survey experiment do recover the corresponding descriptive parameters in Swiss citizenship elections (Hainmueller et al. 2015; see also Auerbach and Thachil 2018). Moreover, when applied to candidate or party choice, the AMCE can be interpreted as the increase in the average vote share attributed to the presence of a specific attribute level (Bansak et al. 2020). In sum, the AMCE is a highly useful and informative summary of aggregate preferences in multidimensional settings.

That said, it does not mean that any average causal effect necessarily tells the whole story. Just as an ATE can hide important heterogeneity in the individual-level causal effects, the AMCE might also hide such heterogeneity – for example, if the effect of an attribute value is negative for one half of the sample and positive for the other half. In such settings, conditional AMCEs for relevant subgroups might be useful to explore, as we discuss later in this section. Similarly, just as a positive ATE does not necessarily imply that a treatment has positive individual-level effects for a majority of subjects, a positive AMCE does not imply that a majority of respondents prefer the

attribute value in question (Abramson et al. 2019). Researchers should accordingly be careful in their choice of language for describing the substantive interpretations of AMCEs as an average causal effect.

2.5.3 More on Estimation and Inference

Despite the high dimensionality of the design matrix for our factorial conjoint treatments, the AMCEs in our Democratic primary experiment are reasonably precisely estimated based on 503 respondents, as can be seen from the widths of the confidence intervals in Figure 2.3. Many applied survey experimentalists find this rather surprising, since it seems to run counter to the conventional wisdom of being conservative in adding treatments to factorial experiments. What is the "trick" behind this?

The answer to this question lies in the implicit averaging of the profile-specific treatment effects in the definition of the AMCE. Once we focus on a particular attribute of interest, the remaining attributes become covariates (that also happen to be randomly assigned) for the purpose of

estimating the particular AMCE. This implies that those attributes simply add to the infinite list of pretreatment covariates that might also vary across respondents or tasks, which are also implicitly averaged over when calculating the observed difference in means. Thus, a valid inference can be made for the AMCE by simply treating the attribute of interest as if it was the sole categorical treatment in the experiment, although statistical efficiency might be improved by explicitly incorporating the other attributes in the analysis.

A straightforward method to incorporate information about all of the attributes in estimating the individual AMCEs for the sake of efficiency is to run a linear regression of the observed outcome on the entire set of attributes, each being "dummied out" with the baseline value set as the omitted category. The estimates presented in Figure 2.3 are based on this methodology instead of individual differences in means. The multiple regression approach has the added benefit of the convenience that one can estimate the AMCEs for all attributes at once, and despite the superficial use of a linear regression model, it requires no functional form assumption by virtue of full randomization.[5] Thus, this approach is currently the most popular in applied studies.

These estimation methods can be applied to various types of outcome variables – such as binary choices, rankings, and ratings – without modification. For illustration, Figure 2.4 shows estimated AMCEs for each of the 10 attributes included in our Democratic primary experiment, along with their 95% confidence intervals, using the seven-point scale rating instead of the forced choice item as the outcome measure. In this application, the estimated AMCEs from the rating outcome are similar to those from the forced choice outcome. Such a similar pattern between the two types of outcomes is frequently, but not always, observed in our experience with conjoint experiments.
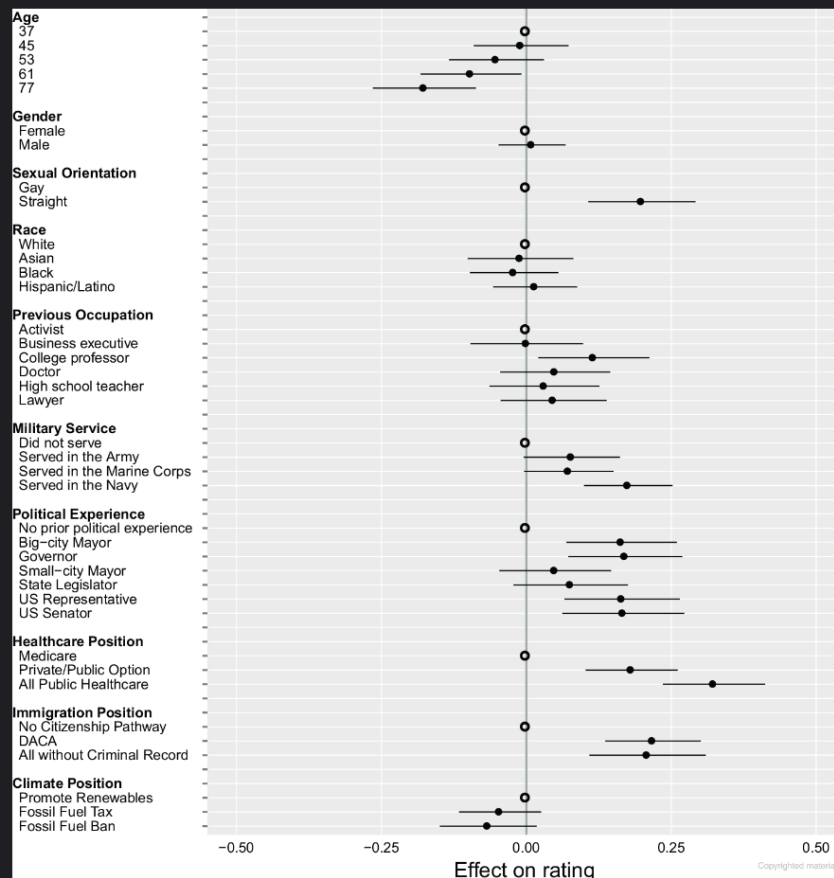
Figure 2.4 Average marginal component effects of candidate attributes in the Democratic primary conjoint experiment (rating outcome).

2.5.4 Conditional AMCE

Another common quantity of interest in conjoint applications in political science is the conditional AMCE, or the AMCE for a particular subgroup of respondents defined based on a pretreatment

respondent characteristic (Hainmueller et al. 2014; also see Chapter 13 in this volume). In our Democratic primary experiment, a natural question of substantive interest is whether preferences about hypothetical Democratic nominees might differ depending on respondents' partisanship. To answer this question, we analyze the conditional AMCEs of the attributes by estimating the effects for different respondent subgroups based on their partisanship.

Figure 2.5 shows the estimated conditional AMCEs for Democratic, independent, and Republican respondents, respectively. As we anticipated, the AMCEs for the policy position attributes are highly variable depending on whether a respondent is a Democrat or a Republican. For example, among Democrats, the probability of supporting a candidate increases by 19 percentage points on average when the position on healthcare changes from supporting Medicare to supporting government healthcare for all. There is no such effect among Republican respondents. There is a similar asymmetry for the effect of the position on immigration. Among Democrats, the probability of supporting a candidate increases by 18 percentage points on

average when the position on immigration changes from supporting no pathway to citizenship for undocumented immigrants to supporting a pathway for all undocumented immigrants without a criminal record. Among Republicans, a similar change in the candidate's immigration position leads to a 11 percentage point decrease in support. Respondents of different partisanship also exhibit preferences in line with their distinct electoral contexts. For example, while prior political experience of the candidate increases support among Democratic respondents on average compared to candidates with no experience in elected office, there is no such effect among Republican respondents.

Figure 2.5 Conditional average marginal component effects of candidate attributes across respondent party.

In interpreting conditional AMCEs, researchers should keep in mind the same set of important nuances and common pitfalls as they do when analyzing AMCEs. That is, they represent an average effect of an attribute level against a particular baseline level of the same attribute, given a particular randomization distribution. In addition, researchers need to exercise caution when comparing a conditional AMCE against another. This is because the difference between two conditional AMCEs does not generally represent a causal effect of the conditioning respondent-level variable, unless the variable itself was also randomly assigned by the researcher. For example, in the Democratic primary experiment, the AMCE of a candidate supporting government healthcare for all as opposed to Medicare was 18 percentage points larger for Democrats than for Republican respondents, but it would be incorrect to describe this difference as a causal effect of partisanship on respondents' preferences for all public healthcare. This point is, of course, no different from the usual advice for interpreting
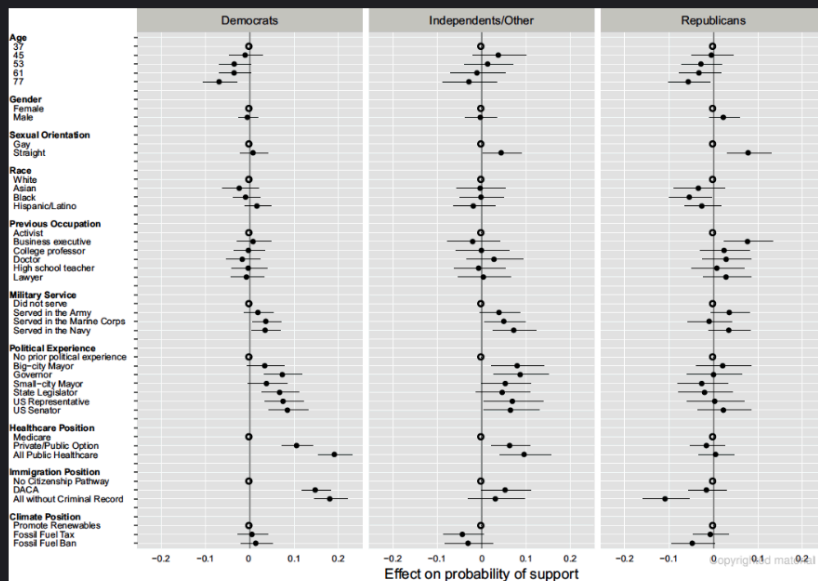
heterogeneous causal effects (such as conditional ATEs) when subgroups are defined with respect to nonrandomized pretreatment covariates, though it is often overlooked in interpreting conditional AMCEs in conjoint applications (see Bansak 2020; Leeper et al. 2020).

2.6 Applications of Conjoint Designs in Political Science

As discussed earlier in this chapter, a key factor behind the popularity of conjoint experiments in political science is their close substantive fit with key political science questions. Indeed, conjoint designs have been applied to understand how populations weigh attributes when making various multidimensional political choices, such as voting, assessing immigrants, choosing neighborhoods and housing (Hankinson 2018; Mummolo and Nall 2016), judging climate-related policies (Bechtel et al. 2019; Gampfer et al. 2014; Stokes and Warshaw 2017), publication decisions (Berinsky et al. 2019), and various other problems (Auerbach and Thachil 2018; Ballard-Rosa et al. 2017; Bechtel and Scheve 2013; Bernauer and Nguyen 2015; Gallego and

Marx 2017; Hemker and Rink 2017; Sen 2017).

In Table 2.2, we report the distribution of 124 recent conjoint applications by their broad topical areas published between 2014 and 2019.[6] A plurality of 27% of the applications involve voting and candidate choice. But conjoint designs have been deployed to understand how people collectively weigh different attributes in a wide range of other applications, from politically relevant judgments about individuals to choices among different policy bundles. In the rest of this section, we review several key areas of conjoint applications in more detail.

Table 2.2 Topical classification of the 124 published articles using conjoint designs identified in our literature review for the years 2014–2019.

Topic	Percentage (%)
Voting	27
Public opinion	19
Public policy	6
Immigration	6
Government	6
Climate change	6
Representation	5
International relations	5
Partisanship	4
Other	

2.6.1 Voting

While some classic theoretical models examine political competition over a single dimension (Downs 1957), choosing between real-world candidates and parties almost always requires an assessment of trade-offs in aggregate. Conjoint designs are especially well suited to study how voters make those trade-offs. It is no surprise, then, that candidate and party choice is among the most common applications of conjoint designs (Abrajano et al. 2015; Aguilar et al. 2015; Carnes and Lupu 2016; Crowder-Meyer et al. 2020; Franchino and Zucchini 2015; Horiuchi et al. 2018; Kirkland and Coppock 2018; Teele et al. 2018).

One especially common use of conjoint designs has been to examine biases against candidates who are from potentially disadvantaged categories including women, African Americans, and working-class candidates. Crowder-Meyer et al. (2020), for example, demonstrate that biases against Black candidates increase when MTurk respondents are cognitively taxed. This study also illustrates another advantage of conjoint designs, which is that they permit the straightforward estimation of differences in the causal effects or AMCEs across other randomly assigned variables. Those other variables can either be separate attributes within the conjoint or else randomized interventions external to the conjoint itself. An example of the former would be analyzing the difference in AMCEs across the levels of another randomized attribute, while an example of the latter would be analyzing the difference in AMCEs when the framing of the conjoint task itself varies. As designs of the latter type demonstrate, conjoint analyses can not only provide measures of preferences about multidimensional objects, but also be used to evaluate the effect of separate, randomized interventions (see especially Butler and Crabtree 2017; Dill and Schubiger 2019).

At the same time, conjoint designs can help explain observed biases even when uncovering no outright discrimination. Carnes and Lupu (2016) report conjoint experiments from Britain, the USA, and Argentina showing that voters do not penalize working-class candidates in aggregate, a result that suggests that the shortage of working-class politicians is driven by supply-side factors. Also, Teele et al. (2018) use conjoint designs to

show that American voters and officials do not penalize – and may even collectively favor – female candidates. Yet they also prefer candidates with traditional family roles, setting up a "double bind" for female candidates.[7]

Conjoint designs can also be employed to gauge the associations between attributes and a category of interest (Bansak et al. 2019). For example, Goggin et al. (2019) use conjoint experiments embedded in the Cooperative Congressional Election Study to have respondents guess at candidates' party or ideology using issue priorities and biographical information. They find that low-knowledge and high-knowledge voters alike are able to link issues with parties and ideology, providing grounds for guarded optimism about voters' capacity to link parties with their issue positions. Candidate traits, by contrast, do not provide sufficient information to allow most voters to distinguish the candidates' partisanship. Conjoint designs have thus helped shed new light on long-standing questions of ideology and constraint.

Still other uses of conjoint designs can illuminate aspects of voter decision-making and political psychology. For example, ongoing research by Ryan and Ehlinger (2019) examines a vote choice setup in which candidates takes positions on issues whose importance to the respondents had been identified in a previous wave of a panel survey. And separate research by Bakker et al. (2019) deploys conjoint methods to show that people low in the psychological trait "agreeableness" respond positively to candidates with antiestablishment messages. Conjoint designs can also shed light on how political parties choose which candidates to put before voters in the first place (Doherty et al. 2019). Researchers can also use conjoint designs to examine the interactions among different attributes as well as the effects of specific clusters of attributes.

2.7 Immigration Attitudes

Whether hiring, dating, or just striking up a conversation, people evaluate other people constantly. That may be one reason why conjoint designs evaluating choices about individuals have proven to be relatively straightforward – and often even engaging – for many respondents. Indeed, we commonly find that respondents seem

to enjoy and engage with conjoint surveys, perhaps because of their novelty. In response to one of the experiments done for Bansak et al. (2019), a respondent wrote, "This survey was different than others I have taken. I enjoyed it and it was easy to understand." An MTurk respondent wrote, "Thank you for the fun survey!" Such levels of engagement may help explain some of the robustness of conjoint experiments we detail above.

Given how frequently people find themselves evaluating other people, it is not surprising that conjoint experiments have been used extensively to evaluate immigration attitudes (Adida et al. 2017; Auer et al. 2019; Bansak et al. 2016; Clayton et al. 2019; Flores and Schachter 2018; Hainmueller and Hopkins 2015; Schachter 2016; Wright et al. 2016). Hainmueller and Hopkins (2015) demonstrate that American respondents recruited via a probability sample actually demonstrate surprising agreement on the core attributes that make immigrants to the USA more or less desirable. Wright et al. (2016) show that sizable fractions of American respondents choose not to admit either immigrant when they are presented in pairs and there is the option to reject

both.

2.7.1 Policy Preferences

Another area where conjoint experiments have been employed is to examine voters' policy preferences. In these applications, respondents are often confronted with policy packages that vary on multiple dimensions. Such designs can be used to examine the trade-offs that voters might make between different dimensions of the policy and to examine the impacts of changing the composition of the package. For example, Ballard-Rosa et al. (2017) use a conjoint survey to examine American income tax preferences by presenting respondents with various alternative tax plans that vary the level of taxation across six income brackets. They find that voter opinions are not far from current tax policies, although support for taxing the rich is highly inelastic. Bansak et al. (2021) employ a conjoint experiment to examine mass support in European countries for national austerity packages that vary along multiple types of spending cuts and tax increases, allowing them to evaluate eligible voters' relative sensitivities to different austerity measures as well as to estimate

average levels of support for specific hypothetical packages.

One feature of these studies is that the choice task (i.e., evaluating multidimensional policy packages) is presumably less familiar and more complex for respondents than the task of evaluating people or political candidates. That said, many real-world policies involve precisely this type of multifeatured complexity, and preferences of many voters vis-à-vis these policies might well be highly contingent. For example, respondents might support a Brexit plan only if it is based on a negotiated agreement with the European Union. Similarly, during its 2015 debt crisis, Greece conducted a bailout referendum in which voters were asked to decide whether the country should accept the bailout conditions proposed by the international lenders.

2.7.2 Challenges and Open Questions

Still, there are a range of outstanding questions about conjoint survey experiments. For example, a central challenge in designing conjoint experiments is the possibility of producing unrealistic profiles. Fully randomized conjoint designs have desirable features, but one limitation is that the independent randomization of attributes that are in reality highly correlated may produce profiles that seem highly atypical. To some extent, this is a feature rather than a bug: it is precisely by presenting respondents with atypical profiles that it is possible to disentangle the specific effects of each attribute. While in 2006 it might have seemed unlikely that the next US president would be the son of a White mother from Kansas and a Black father from Kenya and someone who spent time in Indonesia growing up, Barack Obama was inaugurated just a few years later.

In some instances, however, atypical or implausible profiles are genuine problems, which can be addressed through various approaches. For one thing, researchers can modify the incidence of different attributes to reduce the share of profiles that are atypical. They can also place restrictions on attribute combinations or can draw two seemingly separate attributes jointly. For example, if the researchers want to rule out the possibility of a candidate profile of a very liberal Republican, they can simply draw

ideology and partisanship jointly from a set of options that excludes that combination. Finally, researchers can also identify profiles as atypical after the fact and then examine how the AMCEs vary between profiles that are more or less typical, as in Hainmueller and Hopkins (2015). In practice, the atypical profiles are randomly assigned, so researchers can straightforwardly compare AMCEs between profiles or tasks judged to be atypical and others.

There are also outstanding questions about external validity. To date, conjoint designs have been administered primarily via tables with written attribute values, even though information about political candidates or other choices is often processed through visual, aural, or other modes. Do voters, for example, evaluate written attributes presented in a table in the same way that they evaluate attributes presented in more realistic ways? The table-style presentation may prompt respondents to evaluate the choice in different ways, and so hamper external validity. It also has the potential to lead respondents to consider each attribute separately, rather than assessing the profile holistically.

One core benefit of conjoint designs can also be a liability in some instances. Conjoint designs return many possible quantities of interest, allowing researchers to compare the AMCEs for various effects and to test hypotheses competitively. However, this also opens up the possibility of multiple comparisons concerns, as researchers may conduct multiple statistical tests. This feature of conjoint designs makes preregistration and preanalysis plans especially valuable in this context. At the same time, conjoint experiments open up a wide range of new substantive and statistical questions about the interactions across different attributes, questions that researchers have only begun to probe.

References

Abrajano, Marisa A., Christopher S. Elmendorf, and Kevin M. Quinn. 2015. "Using Experiments to Estimate Racially Polarized Voting." UC Davis Legal Studies Research Paper Series, No. 419.

Abramson, Scott F., Korhan Koçak, and Asya Magazinnik. 2019. "What Do We Learn about Voter Preferences from Conjoint Experiments?" Working paper presented at PolMeth XXXVI.

Acharya, Avidit, Matthew Blackwell, and Maya Sen.